AN ADAPTIVE CHOICE OF THE SCALE PARAMETER FOR M-ESTIMATORS

BY

ROBERT MICHAEL BELL

TECHNICAL REPORT NO. 3
JULY 1, 1980

U.S. ARMY RESEARCH OFFICE
RESEARCH TRIANGLE PARK, NORTH CAROLINA
CONTRACT NO. DAAG29-79-C-0166

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

AN ADAPTIVE CHOICE OF THE SCALE PARAMETER FOR M-ESTIMATORS

BY

ROBERT MICHAEL BELL

TECHNICAL REPORT NO. 3
JULY 1, 1980

U.S. ARMY RESEARCH OFFICE
RESEARCH TRIANGLE PARK, NORTH CAROLINA
CONTRACT NO. DAAG29-79-C-0166

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

AN ADAPTIVE CHOICE OF THE SCALE PARAMETER FOR M-ESTIMATORS

Robert Michael Bell, Ph.D.
Stanford University, 1980

Let $x_1,\ldots,x_n$ be a random sample from a distribution symmetric about the unknown location parameter θ. A major class of robust estimators of location is the class of M-estimators, each of which corresponds to a function ψ defined on the reals. To be scale equivariant, these estimators require the use of a scale equivariant function of the sample. Commonly, this scale parameter is chosen to be a constant times the sample MAD (median absolute deviation from the median).

For a given function ψ, the variance of the corresponding M-estimator varies considerably with the value of the scale parameter. It is therefore proposed that the value which minimizes an estimate of the asymptotic variance of the M-estimator be used as the scaling factor. This adaptive method of scaling is shown to be asymptotically optimal (under fairly general conditions), in the sense that the resulting M-estimator has the smallest possible asymptotic variance among all M-estimators based on ψ. In particular, when the underlying distribution is normal, the adaptive estimator based on any reasonable ψ achieves full asymptotic efficiency, i.e., is asymptotically equivalent to the sample mean.

The performance of the estimator for small samples is investigated by Monte Carlo methods for several choices of ψ using the trefficiency criteria. A slight modification of the above estimator compares favorably with Tukey's bisquare M-estimator for sample sizes as small as 20.

ACKNOWLEDGMENTS

I am very grateful to my adviser, Vernon Johns, for his help and encouragement from the birth of the first idea to the writing of the last words. This thesis would not have been possible without his able guidance. I also thank Rupert Miller and Paul Switzer for their comments and criticisms.

I would like to acknowledge Thomas Cover and Persi Diaconis, who have each profoundly influenced my perspective on statistical research. The comments and encouragement of Delores Conway, Michael Cohen, and Barry Eynon are also deeply appreciated.

Finally, I thank Wanda Miller for her expert typing job.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

AN ADAPTIVE CHOICE OF THE SCALE PARAMETER FOR M-ESTIMATORS

1. Introduction

Suppose that one observes a random sample $x_1,\ldots,x_n$ from a distribution symmetric about the unknown location parameter θ. Except in a make-believe world, a desirable property for any estimator of θ is that of robustness. We list two qualitative definitions of robustness of location estimators. The first is that an estimator is robust if it possesses high efficiencies for a wide set of likely distributions. The second requires that the estimator be highly efficient for a particular model and yet resistant to a small amount of contamination. The two concepts are generally compatible, and they are both presented to give alternative motivations for robustness. Certainly, many other definitions can also be given; e.g., see Huber (1972).

A major class of robust estimators is the class of M-estimators of location. In order to be scale equivariant, these estimators require the use of a scale equivariant function of the sample $s(\underset{\sim}{x})$. Commonly $s(\underset{\sim}{x})$ is a constant times the sample MAD (median absolute deviation from the median). In this paper we propose using the value of the scale parameter which minimizes the estimated asymptotic variance of the M-estimator. Under fairly general conditions this converges to the best possible value for the scale parameter, resulting in the smallest possible asymptotic variance for a fixed function ψ. In particular, full asymptotic efficiency is

attained for the normal distribution. In Section 10 we present small sample (n = 20) Monte Carlo results which compare a slight modification of the above proposal with Tukey's bisquare using the trieffi-ciency criteria. Even for this sample size the adaptive estimator compares favorably with the bisquare.

While the analysis in this paper is limited to estimation of location, M-estimation may also be used on the more important problem of multiple linear regression; e.g., see Andrews (1974). There appears to be a straightforward extension of the methods described here to the regression problem.

2. Robust Estimators of Location

The nonrobustness of the sample mean--and other least squares procedures--is well documented. The mean has very low efficiency for long-tailed distributions and gives far too much weight to gross outliers. Robust alternatives to the mean are usually limited to estimators which are location and scale equivariant. In accordance with Berk (1967), an estimator $T(\underset{\sim}{x})$ is location and scale equivariant if $T(a\underset{\sim}{x}+b\underset{\sim}{1}) = aT(\underset{\sim}{x}) + b$, for all $a > 0$. We note that the term invariant is also used by some authors for the above property.

The most commonly used nonadaptive estimators of location fall into three categories:

(i) L-estimators. Linear combinations of order statistics. These include trimmed means which in turn include the mean and median as extremes.

(ii) M-estimators. Analogues of maximum likelihood estimators (MLEs). The M-estimator $\hat{\theta}$ is a root of an equation of the form $\Sigma_{i=1}^{n} \psi[(x_i - \hat{\theta})/s(\underset{\sim}{x})] = 0$, where ψ is a skew-symmetric function.

(iii) R-estimators. Midpoints of symmetric confidence intervals based on linear rank tests. The best known of these is the Hodges-Lehmann (1963) estimator, which turns out to be the median of the pairwise averages of the observations.

Examples of each of these types of estimators are contained in the books by Andrews, et al. (1972) and Huber (1977). Johns (1979) has introduced P-estimators, which are robust analogues of Pitman estimators.

Our further attention will be devoted almost exclusively to M-estimators. The main reason for this is one given by Hampel (1974), "Furthermore, the stress on M-estimates is not accidental; neither L- nor R-estimates...allow a proper rejection of outliers, i.e., a rejection based on the distance from the bulk of the data." This property of M-estimators is closely related to the fact that their influence curves (defined in Section 4) are essentially independent of the true distribution function F. We note that this property is shared by P-estimators. Another reason for preferring M-estimators is that because of their close association to MLEs, they can be generalized to a large number of models.

Just as there are many definitions of robustness, there are many criteria for judging robust estimators. The finite sample variances (or equivalently efficiencies) for various symmetric

distributions are often used. The use of the variance is justified by the argument that symmetry implies unbiasedness and good estimators tend to be approximately normally distributed except when based on small samples from long-tailed distributions. Monte Carlo techniques are usually required to obtain these variances. The most ambitious study of this kind is the Princeton Study by Andrews, et al. (1972). Their study includes 65 estimators and 30 sampling "situations" (distribution and sample size). Sample sizes ranged from 5 to 40.

Except for the obvious choice of the normal, selecting from among the many distributions is difficult. The results presented in this paper are based on Tukey's (1979) concept of triefficiency. These are three diverse sampling situations for which a robust estimator should do well. In each case a sample size of n = 20 is used. The three situations are:

1. Standard normal.

2. One wild normal (1WN): 19 standard normals, and 1 N(0,100) in each sample.

3. Slash: a standard normal random variable divided by an independent uniform (0,1).

The asymptotic variance is often another useful tool. Comparison of the asymptotic variances of similar estimators may give a reasonable idea of their relative performances for even fairly small sample sizes. Care must be taken since the rates of convergence of the finite sample variance to the asymptotic variance may

differ greatly for dissimilar estimators. For example, the convergence for adaptive estimators, defined in the next section, will tend to be slower than that for nonadaptive ones.

Another set of criteria based on resistance to contamination are set forth by Huber (1977) and Hampel (1974). These include breakdown point, maximum bias and variance as a function of the amount of contamination, and gross error sensitivity.

3. Adaptive Estimators

The search for good robust estimators of location often becomes an attempt to find estimators which are efficient for both the normal distribution and longer tailed alternatives like the slash. Unfortunately, tradeoffs must be made. The mean, which is optimal for the normal, does terribly once there is even a moderate amount of contamination. Simple estimators which do very well for long-tailed distributions sacrifice too much efficiency for the normal. Inevitably some compromise must be made between the extreme objectives if we limit ourselves to the nonadaptive estimators described in the last section.

The motivation for adaptive estimators is based on two facts. First, if the approximate shape of the sampling distribution is known, it is not too difficult to find an estimator which will do well by most standards. Second, considerable information about the shape of the sampling distribution is contained in most samples of a moderate size. Adaptive estimators try to use the information in the

sample to select an estimator which is appropriate for the distribution from which the sample came.

In general, adaptive estimators have the following form. A set of estimators, finite or infinite, is chosen. At a minimum the set would include an estimator designed for short-tailed distributions and another for long-tailed ones. Other estimators might be included to handle other sized tails and various degrees of peakedness. Once the set of estimators is given, one needs a choice function which maps the set of possible order statistics into the set of estimators. The objective of the choice function is to choose the estimator which, in some sense, best suits the data.

A good survey article on adaptive estimators is that of Hogg (1974). He proposes a very simple adaptive estimator which chooses from a small number of trimmed means (including a mean of extreme order statistics) based on the value of the Q statistic. The Q statistic is a measure of tail weight based on the outer order statistics. Hogg also suggests that the degree of skewness of the sample may be used to help choose the estimator. Other examples of adaptive estimators are described in Andrews, et al. (1972, Sections 2B3 and 2E1).

An adaptive estimator of particular interest is due to Jaeckel (1971). His estimator is an α-trimmed mean where $\alpha \in [\alpha_0, \alpha_1]$ is chosen adaptively. His method is to use the value of α which minimizes the estimated asymptotic variance of the trimmed mean as a function of α. As $n \to \infty$, the trimming proportion converges in

probability to the optimal trimming proportion in the given range, and thus the asymptotic variance of Jaeckel's estimator is the same as if the optimal trimming proportion were known. He also outlines how this method may be used to choose from among more complex families of L-estimators which are indexed by two or more parameters. Using the estimated variance to choose from a finite set of estimators is also discussed by Switzer (1970).

A more ambitious class of adaptive estimators was introduced by Stein (1956) who argued that estimators could be constructed which would be fully efficient in terms of asymptotic variance for all symmetric densities with finite Fisher information. His argument was that as $n \to \infty$, the density and its derivative can be estimated sufficiently well from the data to produce an estimator with variance approaching the Cramer-Rao lower bound. Several authors have presented estimators with this property. R-estimators were used by Van Eeden (1970) and Beran (1974), while Stone (1975) presented an estimator based on M-estimators. Sequences of L-estimators with efficiencies converging to one have been given by Takeuchi (1971), Johns (1974), and Sacks (1975). Beran (1977) has shown that a minimum Hellinger distance estimator is also fully efficient.

While adaptive estimators are generally designed to have excellent large sample properties, careful consideration must be given to their small sample behavior. Estimators which try to gather too much information from the data will tend to "overadapt." In that case a very wrong inference about the true distribution may

be taken from some samples resulting in a very bad estimate. Thus the amount of adaptation must be carefully pegged to the sample size--with more complex procedures being reserved for larger samples.

For small sample sizes on the order of n = 20, for which only simple adaptive estimators are practical, they have tended to do no better than good nonadaptive estimators. As n increases, the adaptive estimators improve relative to nonadaptive ones. Thus some improvement could be expected for moderate sized samples using some of the fairly simple adaptive estimators. Less can be said about the performance of fully efficient estimators for small and moderate sized samples. The theoretical results for these estimators promise nothing about their performance for finite n, even if n is very large. However, simple estimators from the sequences of Takeuchi and Johns performed well in the Princeton study. Encouraging results are also obtained for n = 40 by Stone (1975).

In this paper we present an adaptive M-estimator for which the function ψ is fixed and the scale parameter $s(\underset{\sim}{x})$ is chosen in the same way as Jaeckel's trimming proportion. This adaptation is simple in the sense that only one parameter is chosen adaptively. In fact, no more parameters than usual are involved in the estimator, since $s(\underset{\sim}{x})$ always depends on the data. Even so, adaptively choosing the scale parameter can lead to substantial improvement. This idea was mentioned as Proposal 3 of Huber (1964), but there appears to have been no previous attempt to pursue it.

4. M-Estimators

M-estimators of location were introduced in 1964 by Peter Huber as analogues of maximum likelihood estimators. For a given function ρ the M-estimator $\hat{\theta}$ is defined as the value of θ which minimizes $\Sigma_{i=1}^{n} \rho[(x_i - \theta)/s(\underset{\sim}{x})]$. The sample scale parameter $s(\underset{\sim}{x})$ is necessary to make $\hat{\theta}$ scale equivariant. Discussion of its importance is postponed until Section 5. It is typically more convenient to define $\hat{\theta}$ as the solution to

$$\sum_{i=1}^{n} \psi\left[\frac{x_i - \hat{\theta}}{s(\underset{\sim}{x})}\right] = 0 \tag{4.1}$$

where $\psi = \rho'$, even though (4.1) may have multiple roots when ρ is not convex. If (4.1) does have multiple roots, then care must be taken to find the correct root. Obviously $\hat{\theta}$ is not changed if ψ is multiplied by a constant.

The influence curve is a powerful tool in the analysis of robust estimation. Suppose that an estimator T is a functional of the empirical distribution function. The influence curve of T evaluated at the distribution F is the function

$$IC_{T,F}(x) = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon}[T((1-\varepsilon)F + \varepsilon\, \delta_x) - T(F)] \tag{4.2}$$

where δ_x is the distribution function of a mass point at x. Essentially $IC_{T,F}(x)$ measures the expected change in $T(F_n)$ from adding a small amount of mass to F at the point x. Hampel (1974) provides an excellent discussion of the relationships between an

estimator's robustness properties and its influence curve.

In general, the influence curve depends on T and F through a complex relationship. For the M-estimator $\hat{\theta}$ given by (4.1) however,

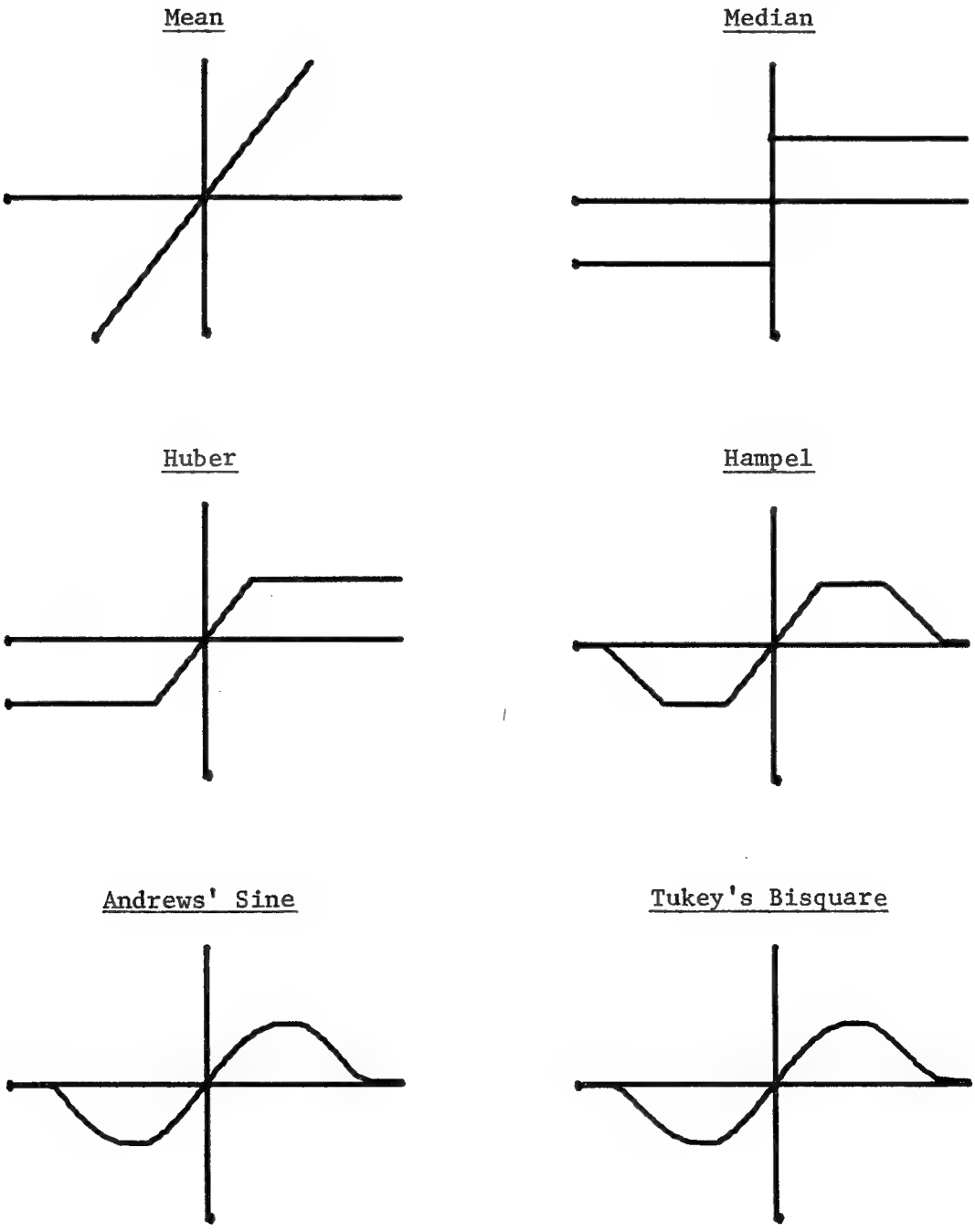
$$IC_{\hat{\theta},F}(x) = \frac{\psi\left(\frac{x-\theta}{x(F)}\right)}{\int \psi'\left(\frac{t-\theta}{s(F)}\right) dF(t)} \tag{4.3}$$

is proportional to ψ at any distribution function F. This fact makes the analysis of M-estimators (and P-estimators, which have a similar expression for $IC_{T,F}(x)$) using the influence curve easier than the analysis of L- or R-estimators. Since $IC_{\hat{\theta},F}(x)$ is so closely related to the function ψ, ψ is often referred to as the influence curve of $\hat{\theta}$. Furthermore, the shape of ψ can be chosen to give $\hat{\theta}$ desirable properties.

The influence curves of six M-estimators are shown in Figure 1. The mean is not robust since its influence curve is unbounded. While the median has a bounded influence curve, it is inefficient at the normal distribution because of the large jump at zero. For an estimator to have good efficiency at the normal, its influence curve should be approximately linear near zero. Huber's ψ combines the best features of the mean and median. Hampel's "redescending" ψ reflects his argument that extreme observations should have very little or no influence. Two other smoothed redescending ψ functions are Andrew's sine and Tukey's bisquare. Huber's ψ is likely to be a good choice if one is interested in the normal

Figure 1

Examples of Influence Curves of M-Estimators



with only a small amount of contamination. A redescending ψ is better if one is also worried about the possibility of very heavy tails.

At this point we list some assumptions which it will be convenient to make about any subsequent functions ψ.

(A1) ψ is skew symmetric; i.e., $\psi(-x) = -\psi(x)$.

(A2) $\psi(x) \geq 0$ for all $x \geq 0$.

(A3) $\psi(x)$ is continuous and piecewise differentiable.

(A4) $\psi'(0) = 1$.

(A5) $\psi''(0) = 0$.

(A6) $\psi'''(0) < 0$.

The sine and bisquare satisfy all the assumptions; but Huber's and Hampel's ψ's fail A6, an assumption which is convenient when the scale is chosen adaptively. Assumption A4 is possible since ψ may be multiplied by a constant without changing $\hat{\theta}$. We list A5 separately even though it is implied by A1 and A6.

Solving (4.1) usually requires an iterative procedure which, of course, must be stopped after a finite number of steps. For the location problem, one or two iterations from a good starting point is often adequate. The one step M-estimator (or m^1-estimator) with starting point M_0 is

$$\hat{\theta}_{(1)} = M_0 + \frac{s(x) \sum_{i=1}^{n} \psi\left(\frac{x_i - M_0}{s(x)}\right)}{\sum_{i=1}^{n} \psi'\left(\frac{x_i - M_0}{s(x)}\right)} \quad . \tag{4.4}$$

The usual starting point is the median.

5. The Scale Parameter

The scale parameter, the function of the data required to make $\hat{\theta}$ scale equivariant, must satisfy $s(a\underset{\sim}{x}+b\underset{\sim}{1}) = a \cdot s(\underset{\sim}{x})$, for $a > 0$. While usually given little attention in previous studies of M-estimators, the scale parameter is key to the performance of the estimator. Simply stated, when $s(\underset{\sim}{x})$ is small, $(x_i - \hat{\theta})/s(\underset{\sim}{x})$ is more likely to be on the wings of ψ where little or no weight is given the observations. The result is a very outlier resistant estimator. When $s(\underset{\sim}{x})$ is large, $(x_i - \hat{\theta})/s(\underset{\sim}{x})$ tends to fall on the approximately linear part of ψ where the observations receive nearly equal weights. Most commonly, $s(\underset{\sim}{x})$ is a constant times the median absolute deviation from the median (MAD) given by

$$\text{MAD} = \text{median}\,\{|x_i - \text{med}(\underset{\sim}{x})|\} \quad . \tag{5.1}$$

In this case the constant multiplying the MAD must compromise between the objectives for light and heavy tails. The same compromise is necessary for MLE type scale parameters solving $\Sigma_{i=1}^{n} \gamma[(x_i - \hat{\theta})/s(\underset{\sim}{x})] = 0$.

As $s(\underset{\sim}{x}) \to \infty$, the estimate $\hat{\theta}(\underset{\sim}{x})$ converges to the sample mean. Because the mean is optimal for the normal, it is convenient to have it as a finite point in the set of estimators. This is accomplished by letting

$$\lambda(\underset{\sim}{x}) = 1/s(\underset{\sim}{x}) \quad , \tag{5.2}$$

so that $\hat{\theta}$ converges to the mean as $\lambda(\underset{\sim}{x}) \to 0$. We will refer to λ as the scale factor. For doing analysis, the resulting parameterization appears to be more natural. The equations defining $\hat{\theta}$ and $\hat{\theta}_{(1)}$ become

$$\sum_{i=1}^{n} \psi[\lambda(x_i - \hat{\theta})] = 0 \tag{5.3}$$

and

$$\hat{\theta}_{(1)} = M_0 + \frac{\sum \psi[\lambda(x_i - M_0)]}{\lambda \sum \psi'[\lambda(x_i - M_0)]} \tag{5.4}$$

for $\lambda = \lambda(\underset{\sim}{x}) > 0$. In each case, the limit as $\lambda \to 0$ is $\bar{x}$. The estimators are equivariant if and only if $\lambda(a\underset{\sim}{x} + b\underset{\sim}{1}) = \lambda(\underset{\sim}{x})/a$, for $a > 0$.

6. Asymptotic Variance of M-Estimators

Under fairly mild regularity conditions on the true distribution, most robust estimators of location have a limiting normal distribution of the form

$$\sqrt{n}\,(T(\underset{\sim}{x}) - \theta) \xrightarrow{\mathcal{D}} N(0, V(T,F)) \quad . \tag{6.1}$$

The variance of the limiting distribution, $V(T,F)$, is called the asymptotic variance of T. Suppose that $\lambda(\underset{\sim}{x}) = \lambda + O_p(n^{-\frac{1}{2}})$ as $n \to \infty$; then for a given function ψ the asymptotic variance of the M-estimator $\hat{\theta}$ is

$$V(\lambda; \psi, F) = \frac{E\,\psi[\lambda(X - \theta)]^2}{\lambda^2 [E\,\psi'[\lambda(X - \theta)]]^2} \quad . \tag{6.2}$$

The asymptotic variance of the one-step M-estimator $\hat{\theta}_{(1)}$ is also $V(\lambda; \psi, F)$.

In Section 5 we gave a heuristic argument to suggest that for fixed ψ, small values of λ are good for short tailed distributions

while larger values of λ are best for long tailed ones. Figure 2 shows the asymptotic variance of the bisquare as a function of λ for seven symmetric densities. The densities have been scaled to have the same MAD, and the variances are normalized by dividing by the Cramer-Rao lower bound, $[\int(f'(x)^2/f(x))dx]^{-1}$. For all except the most peaked densities (Laplace and Cauchy), there exists some λ such that the variance is 1.04 or less; the minimum for the Cauchy is 1.14. However, the points where the minima are achieved are well dispersed. For the normal the minimum occurs at $\lambda = 0$, and as the weight of the tails increase the optimal value of λ does also. If a compromise value of λ is chosen, then at least 5 to 10 percent efficiency must be sacrificed for both the normal and the slash. We note that for each F, $V(\lambda; \psi,F) \to \infty$ as $\lambda \to \infty$. This is because the bisquare redescends to zero allowing only a fraction of the data to be used. As $\lambda \to \infty$, this fraction converges to zero.

Figure 3 compares asymptotic and finite sample (n = 20) variances for three distributions. The contaminated normal for which the asymptotic variances are given in Figures 2 and 3 is 95% N(0,1) and 5% N(0,100). The finite sample variance is for the one wild normal--19 points N(0,1) and 1 point N(0,100). Except for the normal distribution, the Cramer-Rao lower bound is generally not attainable for finite sample sizes. For the slash, the sample size 20 variances are up to 15 percent higher than the corresponding asymptotic variances. Even so, the shapes of the curves are similar, illustrating that comparable sacrifices must be made in choosing a single value of λ.

Figure 2

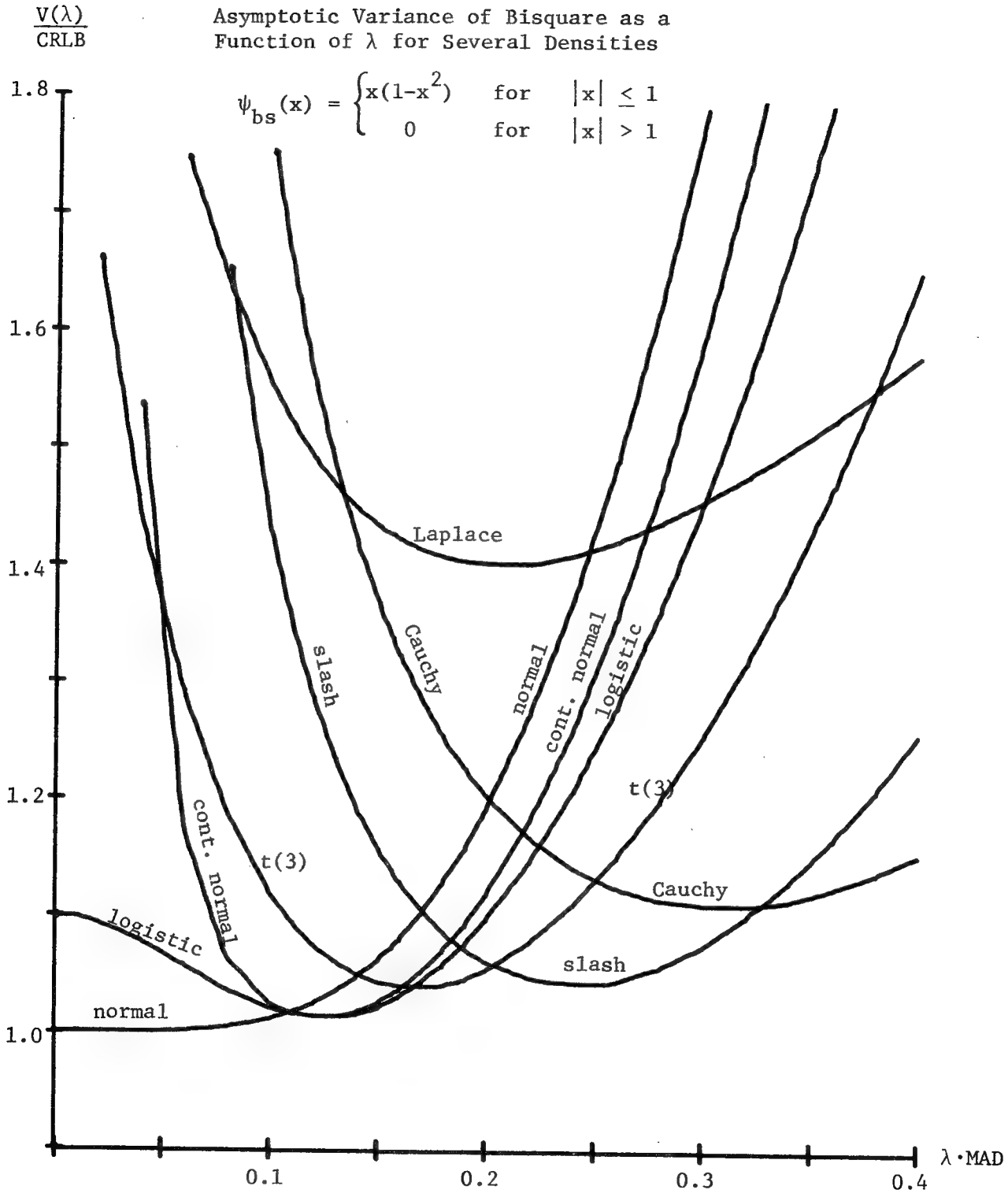
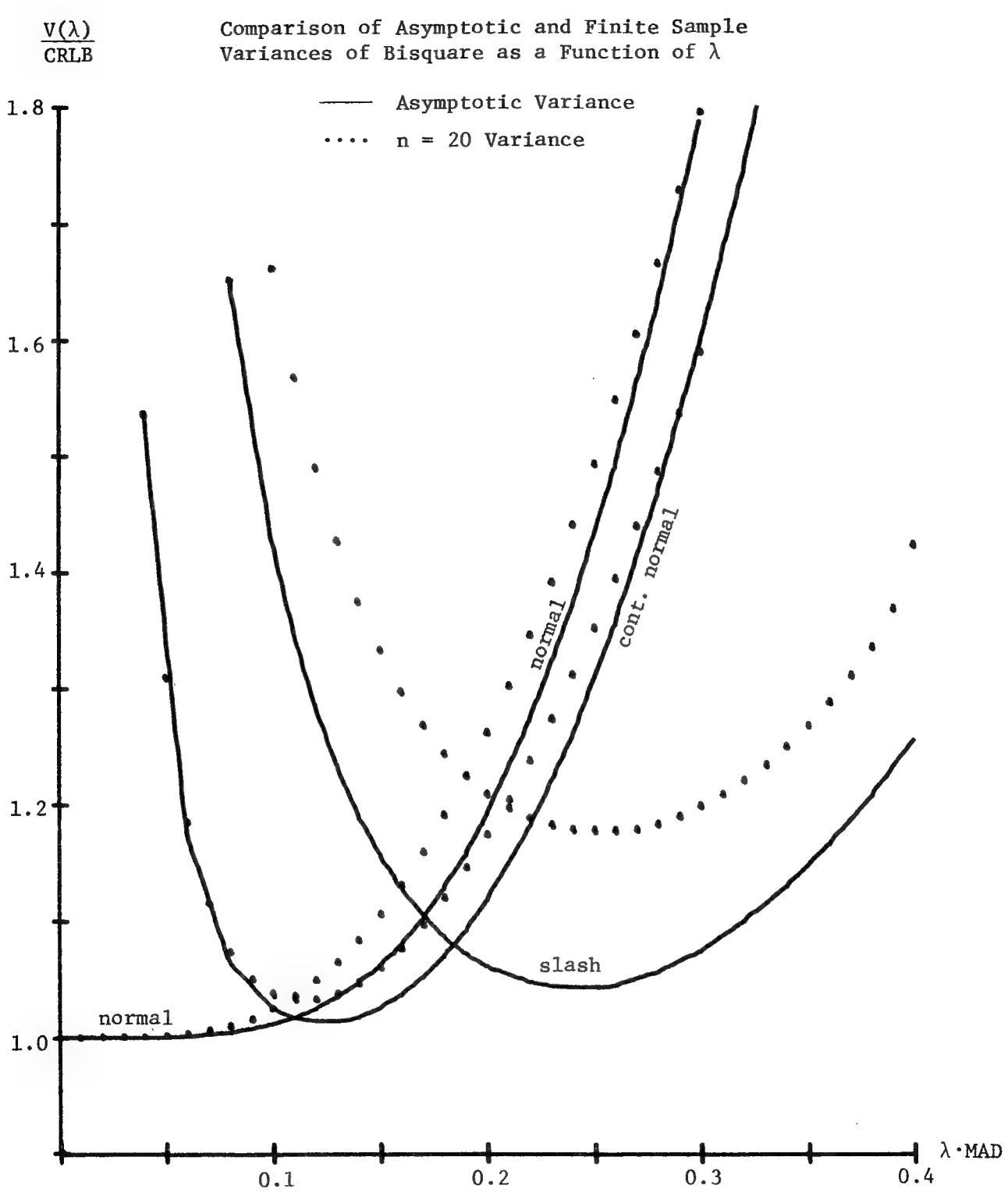

V(λ)/CRLB
Asymptotic Variance of Bisquare as a Function of λ for Several Densities
$\psi_{bs}(x) = \begin{cases} x(1-x^2) & \text{for } |x| \le 1 \\ 0 & \text{for } |x| > 1 \end{cases}$
1.8
1.6
1.4
1.2
1.0
Laplace
slash
Cauchy
normal
cont. normal
logistic
t(3)
cont. normal
t(3)
logistic
normal
slash
Cauchy
0.1
0.2
0.3
0.4
λ·MAD

Figure 3


V(λ)
CRLB
Comparison of Asymptotic and Finite Sample
Variances of Bisquare as a Function of λ
Asymptotic Variance
n = 20 Variance
1.8
1.6
1.4
1.2
1.0
normal
cont. normal
slash
normal
0.1
0.2
0.3
0.4
λ·MAD

Because of the sharp peak in the Laplace density, the bisquare does very poorly for all values of λ. Someone wishing to do well for the Laplace would probably choose a function ψ which is monotone. Figure 4 shows the asymptotic variance of Huber's estimator as a function of λ for the same densities as in Figure 2. For Huber's ψ, $\hat{\theta}$ converges to the sample median as $\lambda \to \infty$. Thus $\hat{\theta}$ becomes efficient for the Laplace as $\lambda \to \infty$. This improvement comes at the expense of increased optimal variances for the contaminated normal (1.11) and slash (1.12).

7. Adaptive Choice of λ

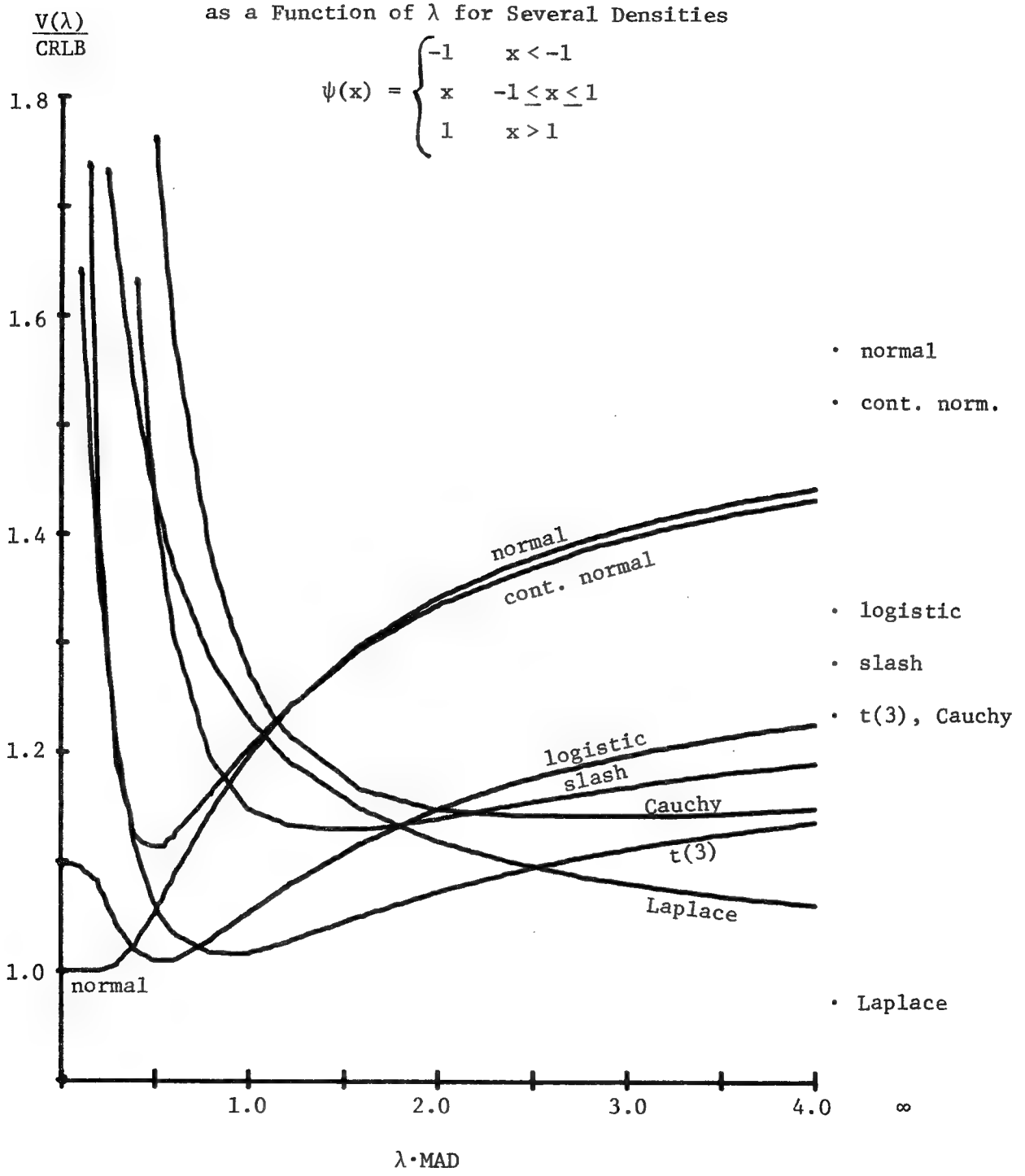
The graphs in the previous section show that for a given function ψ, the choice of λ makes a considerable difference in the variance of the M-estimator. If we could always use λ equal to or nearly equal to the value which minimizes $V(\lambda) = V(\lambda; \psi, F)$, say λ_0, the resulting estimator would do exceptionally well. This suggests trying to estimate λ_0 from the data. The simplest idea is to estimate the function $V(\lambda)$ by a function $\hat{V}(\lambda)$ depending on the data and use the value of λ which minimizes $\hat{V}(\lambda)$, in either (5.3) or (5.4). We denote this value of λ by $\lambda^*(\underset{\sim}{x})$, or just λ^*, and the resulting estimators by $\hat{\theta}^*$ and $\hat{\theta}^*_{(1)}$. Although we shall find it necessary to make a minor modification in this definition of λ^*, the flavor will remain the same.

This procedure is promising for two reasons. First, under most conditions of interest λ^* will converge to λ_0 as $n \to \infty$, and the

Figure 4

Asymptotic Variance of Huber's Estimator as a Function of λ for Several Densities

$$\psi(x) = \begin{cases} -1 & x < -1 \\ x & -1 \le x \le 1 \\ 1 & x > 1 \end{cases}$$



asymptotic variance of the adaptive estimator will be $V(\lambda_0)$, (see Theorems 2-7 in Section 9). If ψ is the bisquare, then the asymptotic efficiency will be 0.96 or higher for five of the seven distributions in Figure 2. The asymptotic efficiency for the normal will be one. Second, the fact that only one parameter is being estimated from the data means that $\hat{\theta}^*$ should be stable for even fairly small n. Working with the scale factor essentially allows us to change the shape of the influence curve adaptively without adapting ψ. The influence curve transforms smoothly from nearly linear (small λ) to quickly redescending (large λ).

For skew-symmetric ψ the asymptotic variance $V(\lambda; \psi, F)$ depends on X only through the distribution of $|X-\theta|$. If we assume that F is symmetric, it is reasonable to let the estimated variance $\hat{V}(\lambda)$ depend only on

$$y_i = |x_i - M| \quad \text{for} \quad i = 1,2,\ldots,n \tag{7.1}$$

where M is an estimate of θ. So that it may be computed directly, we have let M equal the sample median. Other robust estimates could also be used. A choice of particular interest is the M-estimate solving $\Sigma_{i=1}^{n} \psi[\lambda(x_i - M)] = 0$. Unfortunately, since it depends on λ, the calculation of λ^* is made more difficult. Replacing $(X-\theta)$ by y_i and integrals by summations in (6.2), we get

$$\hat{V}(\lambda) = \frac{n \sum_{i=1}^{n} \psi(\lambda y_i)^2}{\lambda^2 \left[\sum_{i=1}^{n} \psi'(\lambda y_i)\right]^2} \; . \tag{7.2}$$

Figures 5-7 show $\hat{V}(\lambda)$ for samples of size 20 from each of the triefficiency distributions. In each case ψ is the bisquare, and the range of λ (in terms of the sample MAD) is twice that of Figure 2. Starting with the normal distribution in Figure 5, the first observation to make is that $\hat{V}(\lambda)$ often has multiple relative minima. This is caused by the extreme instability of $\hat{V}(\lambda)$ as λ increases. The six graphs fairly well represent the likely patterns of $\hat{V}(\lambda)$. After having a relative minimum (usually at or near $\lambda = 0$), the estimated variance may rocket to unreasonable heights (in the hundreds or thousands) before plummeting, often below its original minimum. The volatility of $\hat{V}(\lambda)$ is caused mainly by the large relative changes which can occur in $\Sigma\ \psi'(\lambda y_i)$ as that sum becomes small. Similar patterns appear in the graphs for the 1WN and slash. It is clear that past the first relative minimum the estimated variance in small samples cannot be trusted. Thus any subsequent relative or absolute minima must be considered spurious.

In contrast, the behavior of $\hat{V}(\lambda)$ up to the point of the first relative minimum is usually quite reasonable. This is because $\hat{V}(\lambda)$ generally has a relative minimum before $\Sigma\ \psi'(\lambda y_i)$ becomes small enough to make $\hat{V}(\lambda)$ unreliable. This is not to say that $\hat{V}(\lambda)$ will be an accurate estimate of $V(\lambda)$ for n as small as 20. Fortunately, that is unnecessary. What is necessary is for the shapes of $\hat{V}(\lambda)$ and $V(\lambda)$ to be similar enough so that λ^* adequately approximates λ_0.

These considerations lead us to modify the original proposal and to let λ^* be the smallest $\lambda \geq 0$ such that $\hat{V}(\lambda)$ has a relative

Figure 5

Graphs of $\hat{V}(\lambda)$ Versus $\lambda \cdot MAD$ Using ψ_{bs} for Random Samples of Size 20 From the Normal Distribution

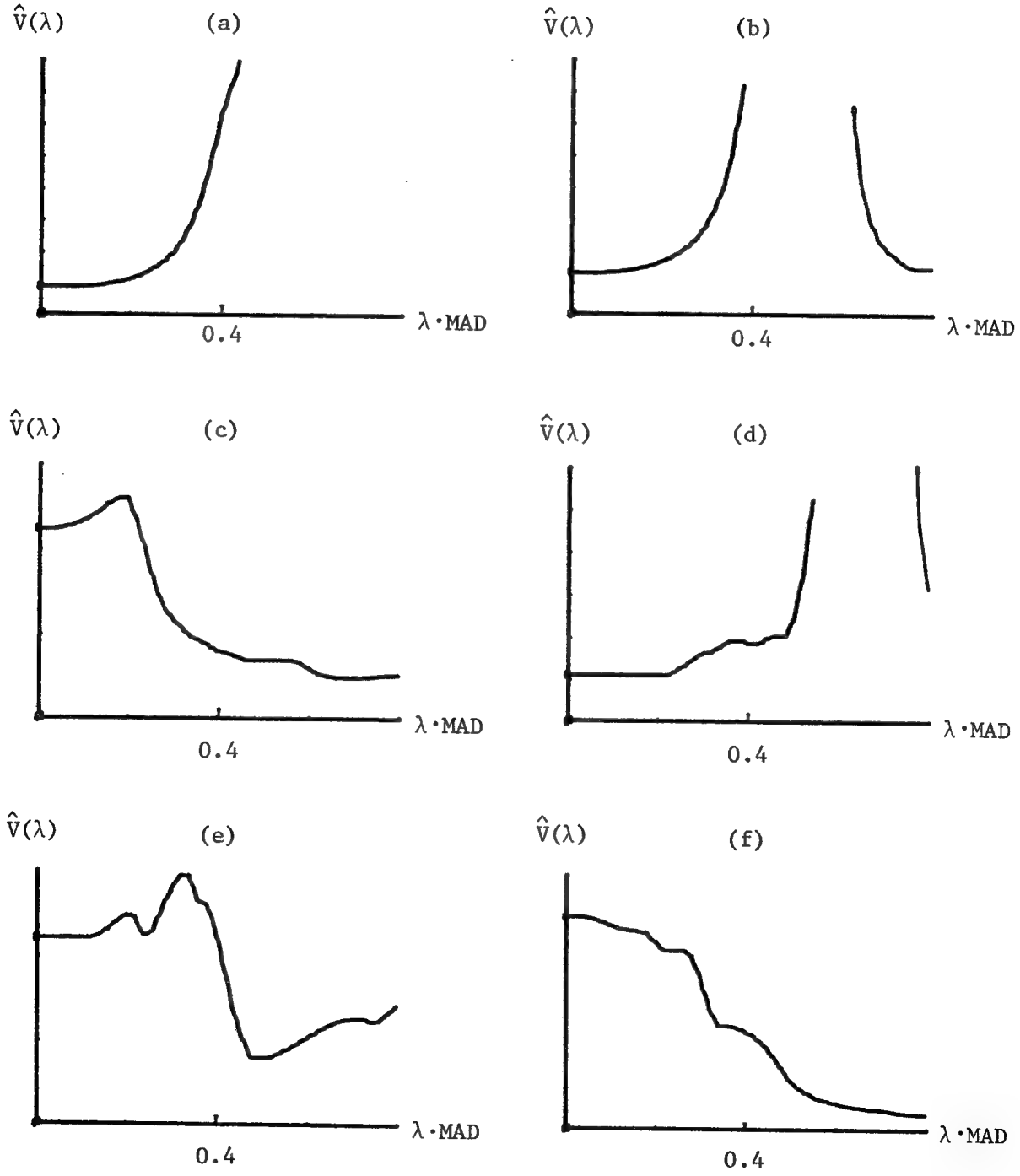


Figure 6

Graphs of $\hat{V}(\lambda)$ Versus $\lambda \cdot MAD$ Using ψ_{bs} for Random Samples of Size 20 From the One Wild Normal

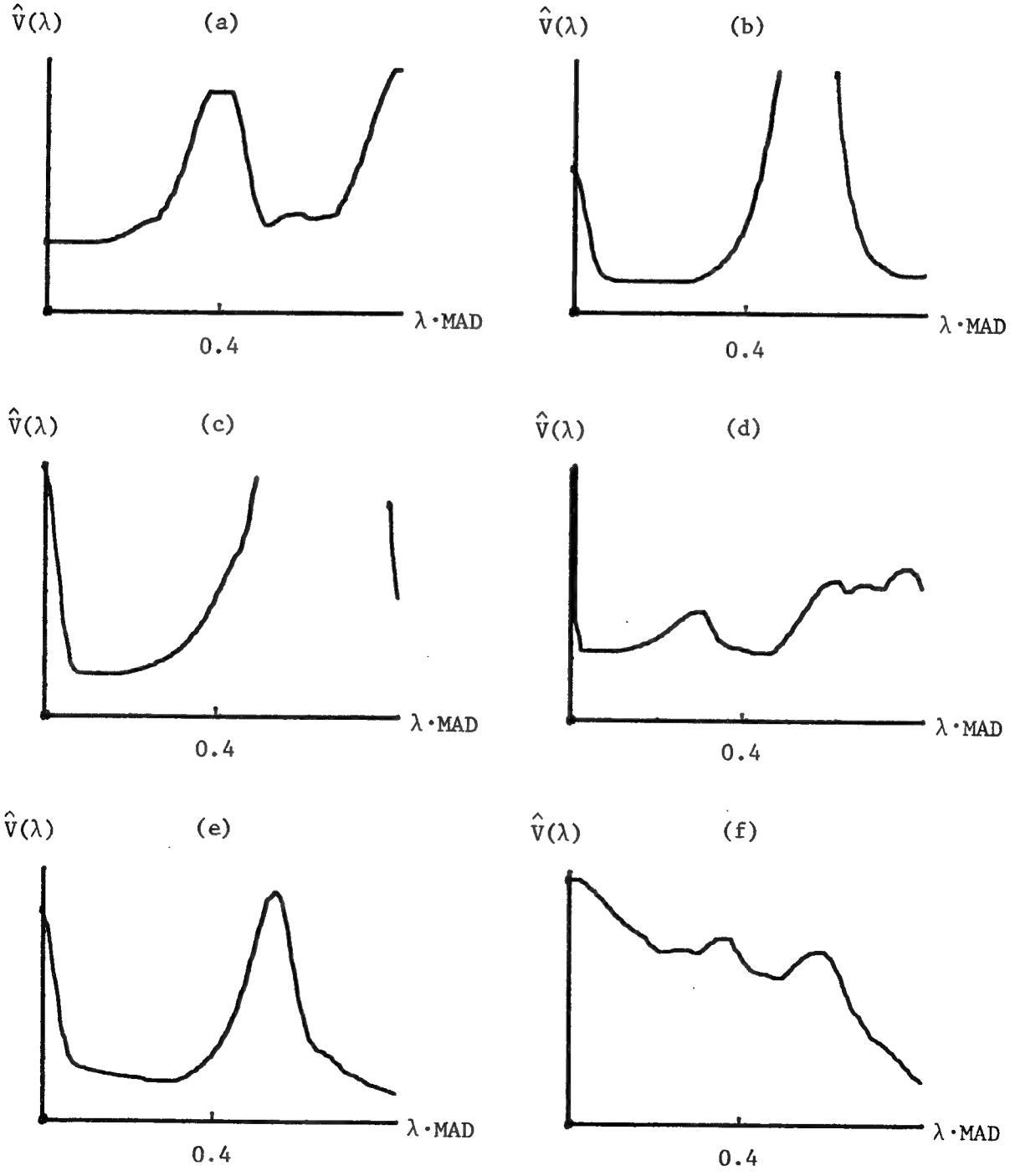
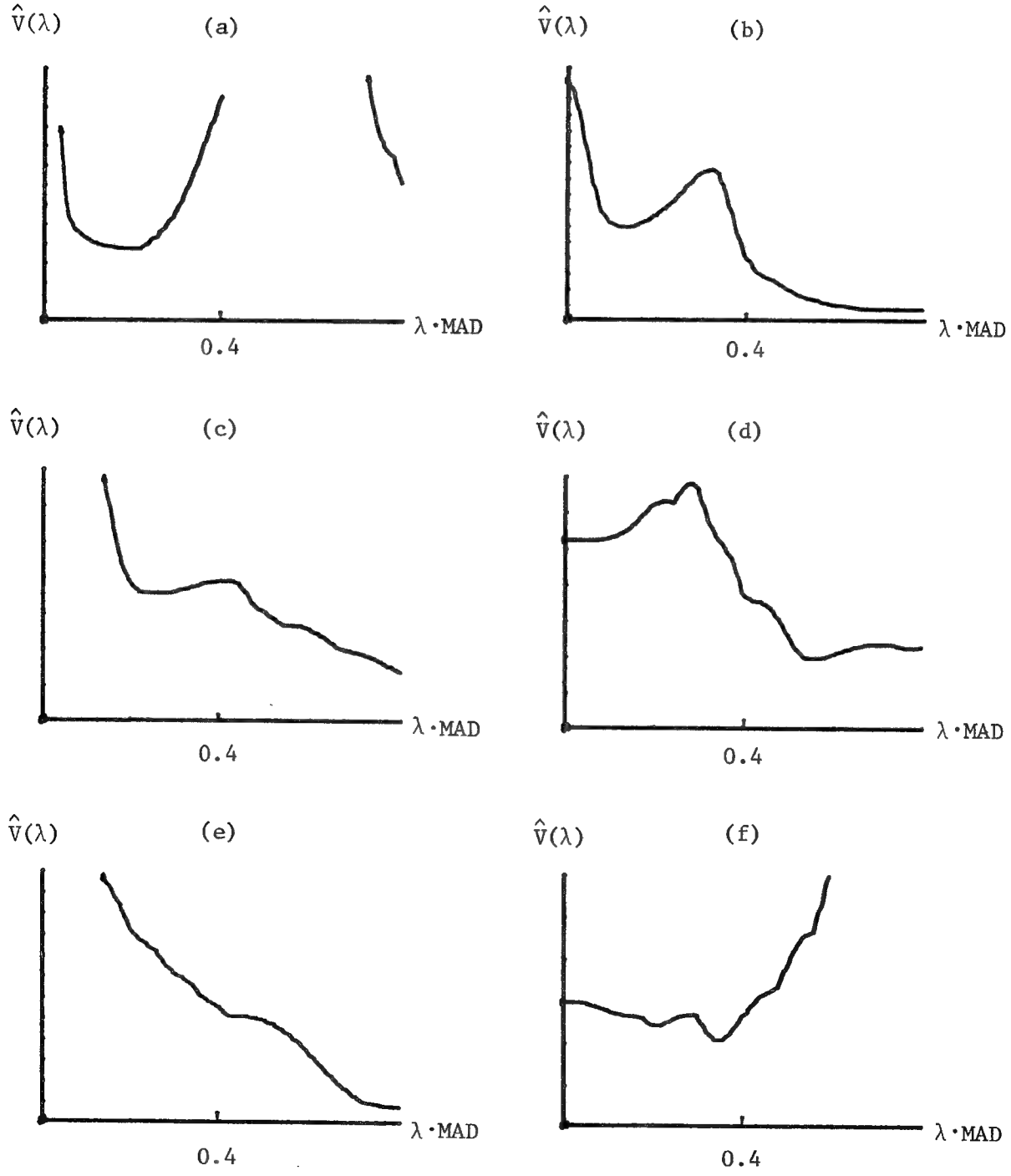


Figure 7

Graphs of $\hat{V}(\lambda)$ Versus $\lambda \cdot MAD$ Using ψ_{bs} for Random Samples of Size 20 From the Slash Distribution



minimum at λ^*. We note that λ^* must exist for redescending ψ since $\Sigma\ \psi(\lambda y_i)$ eventually goes to zero, at which point $\hat{V}(\lambda)$ goes to $+\infty$. Examination of the graphs in Figures 5-7 shows that there is sometimes a relative minimum at $\lambda = 0$. In those cases $\hat{\theta}^*$ is the sample mean. In all other cases $\hat{V}(\lambda)$ has a relative maximum at $\lambda = 0$. This fact is verified by expanding $\hat{V}(\lambda)$ in a Taylor series about $\lambda = 0$. Using A1-A6 on equation (7.2) gives

$$(7.3)\qquad \hat{V}(\lambda) = \frac{n\ \Sigma[\lambda y_i + \frac{1}{6}\lambda^3\ y_i^3\psi'''(0) + o(\lambda^3)]^2}{\lambda^2[\Sigma(1 + \frac{1}{2}\lambda^2\ y_i^2\psi'''(0) + o(\lambda^2))]^2}$$

$$= \frac{1}{n}\Sigma\ y_i^2 + \lambda^2\left[\frac{1}{3n}\Sigma\ y_i^4 - (\frac{1}{n}\Sigma\ y_i^2)^2\right]\psi'''(0) + o(\lambda^2) \quad .$$

Thus we have

$$(7.4)\qquad \hat{V}(0) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - M)^2 \quad ,$$

$$(7.5)\qquad V'(0) = 0 \quad ,$$

and

$$(7.6)\qquad \hat{V}''(0) = \frac{2}{3}\psi'''(0)\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - M)^4 - 3\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - M)^2\right)^2\right] \quad .$$

Since $\hat{V}'(0) = 0$, there is a relative minimum at $\lambda = 0$ (except for a set of measure zero) if and only if $\hat{V}''(0) > 0$. Since $\psi'''(0) < 0$ by A6, $\lambda^* = 0$ if and only if $K(\underset{\sim}{x}) = \dfrac{\frac{1}{n}\Sigma(x_i - M)^4}{[\frac{1}{n}\Sigma(x_i - M)^2]^2} - 3 < 0$. The quantity $K(\underset{\sim}{x})$ might be referred to as the sample kurtosis about

the median. In samples of size 20 from a normal distribution, $K(\underset{\sim}{x})$ is negative about 67 percent of the time. Thus two-thirds of the time $\hat{\theta}^*$ is simply the sample mean. From Figures 6a and 7d it can be seen that $\lambda^* = 0$ sometimes for samples from the 1WN (16 percent of samples) and slash (4 percent) as well. While this is expected for the 1WN, it is at first alarming for the slash. Closer inspection shows that conditional on samples from the slash with $K(\underset{\sim}{x}) < 0$, the sample mean is certainly worse than more resistant estimators, but not disastrously so.

If $\hat{V}'(\lambda)$ is piecewise continuous, then

$$\lambda^* = \inf\{\lambda > 0 : \hat{V}'(\lambda) > 0\} \tag{7.7}$$

where

$$\hat{V}'(\lambda) = \frac{2n}{\lambda^3[\Sigma\, \psi'(\lambda y_i)]^2}\left[\Sigma\, \lambda y_i \psi(\lambda y_i)\psi'(\lambda y_i) - \Sigma\, \psi(\lambda y_i)^2 - \frac{\Sigma\, \psi(\lambda y_i)^2\, \Sigma\, \lambda y_i \psi''(\lambda y_i)}{\Sigma\, \psi'(\lambda y_i)}\right] . \tag{7.8}$$

Of course, the positive factor before the brackets in (7.8) is unnecessary in defining λ^*. The function $\hat{V}'(\lambda)$ depends on the second derivative of ψ. Because of the discontinuity of ψ''_{bs} at $x = 1$, $\hat{V}'(\lambda)$ has positive jumps at $\lambda = 1/y_i$, for $i = 1,2,\ldots,n$. These jumps often cause λ^* to equal $1/y_i$ for one of the larger values of y_i even when the general trend of $\hat{V}(\lambda)$ is downward in that neighborhood. The result is that $\hat{\theta}^*$ does much worse for the long tailed slash than it

would if ψ were smoother. For that reason, as well as for mathematical considerations, we will limit further considerations to functions ψ with several derivatives everywhere. There does not appear to be any practical loss in doing so. The infinitely differentiable functions used in the Monte Carlo study are

$$\psi_p(x) = x\left[1 + \frac{x^2}{2p-1}\right]^{-p} \quad \text{for} \quad p > \frac{1}{2} \tag{7.9}$$

and

$$\psi_\infty(x) = \lim_{p\to\infty} \psi_p(x) = xe^{-(x^2/2)} \quad . \tag{7.10}$$

The maximum of ψ_p always occurs at $x = 1$. As p increases, ψ_p redescends more quickly with ψ_∞ resembling the bisquare (see Figure 8). Monte Carlo results are reported for $p = 1.5$, 2.0, 3.0, and ∞.

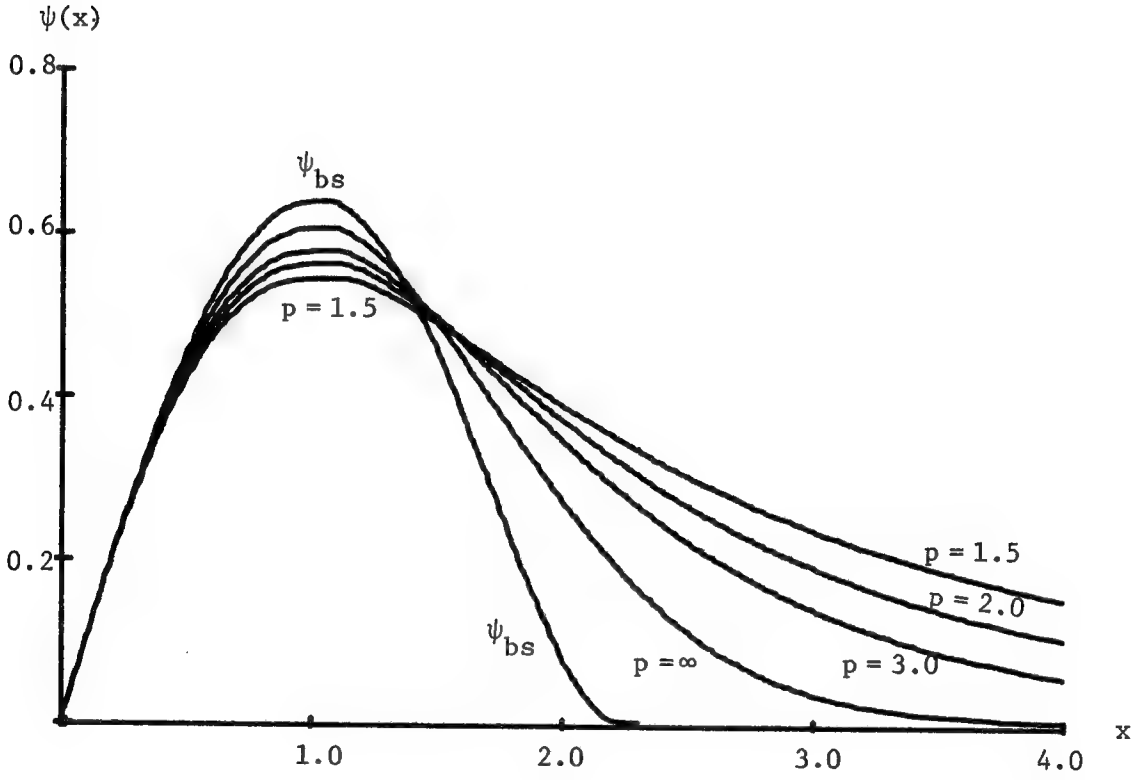
A simple algorithm is sufficient to compute λ^*:

1. Check if $K(\underset{\sim}{x}) < 0$; if so, $\lambda^* = 0$ and $\hat{\theta}^* = \bar{x}$.
2. If not, calculate $\hat{V}'(\lambda)$ for $\lambda = y_{[n]}^{-1}, y_{[n-1]}^{-1}, \ldots$ until it becomes positive.
3. Do a binary search between the last two values of λ until λ^* is known within acceptable tolerance limits.
4. Approximate λ^* by linear interpolation.

Although steps 2 and 3 are not guaranteed to find the first time $\hat{V}'(\lambda)$ becomes positive, the algorithm rarely misses λ^* by much. In fact, for large n the cost might be cut safely by skipping some of the

Figure 8

Graphs of ψ_p for p = 1.5, 2.0, 3.0, ∞ and of ψ_{bs}



evaluations in Step 2. Because of the complexity of $\hat{V}''(\lambda)$ relative to $\hat{V}'(\lambda)$, gradient methods do not seem to offer much of an improvement over the above algorithm. The listing of a FORTRAN program which finds λ^* (with a modification discussed in the next section) and $\hat{\theta}^*_{(1)}$ appears in Appendix A.

8. Modification of λ^*

The choice of λ^* defined by (7.7) leads to an M-estimator which performs poorly on small samples from short tailed distributions. For $n = 20$ the absolute efficiencies of $\hat{\theta}^*_{(1)}$ on the normal and 1WN are at most 0.85 and 0.80, respectively, while we would like each to be at least 0.90. Conversely, the efficiency on the slash is excellent. This evidence suggests that λ^* tends to be too large on average. The simplest attempt to correct this problem, replacing λ^* with $c\lambda^*$ for some $c < 1$, does not work. While decreasing c improves the performance of $\hat{\theta}^*_{(1)}$ on the normal, the simultaneous loss of efficiency on the slash is too great. Furthermore, no choice of c adds more than 2 or 3 percent to the efficiency on the 1WN.

The graphs of $\hat{V}(\lambda)$ help to explain why the last idea is unsuccessful. Most of the time λ^* is fairly close to λ_0, but occasionally $\hat{V}(\lambda)$ has a continual downward trend very far past λ_0. For ψ_{bs} this phenomenon is illustrated in Figures 5f, 6f, 7c, and 7e. Although a local minimum sometimes occurs before the general downward trend terminates, the presence of such a minimum is very sensitive to small changes in $\hat{V}$. When no local minimum occurs, λ^* is too large to

be satisfactorily corrected by a multiplicative factor without over-compensating on other samples.

This problem is more pronounced when $\psi_{3.0}$ is used. Figure 9 shows $\hat{V}(\lambda; \psi_{3.0})$ for the same 1WN samples used for Figure 6. The horizontal axis has been rescaled to account for the different scaling of $\psi_{3.0}$ and ψ_{bs}. Because $\psi_{3.0}$ is smoother and returns to zero faster than ψ_{bs}, these graphs of $\hat{V}(\lambda)$ are much less volatile than those in Figure 6. In particular, $\hat{V}(\lambda)$ is often very flat for moderate sized λ. While $\lambda^* = 0.15/\text{MAD}$ in 9d, a slight change in $\hat{V}$ could move λ^* to 1.0/MAD or more. It is easier to understand the behavior of λ^* by viewing graphs of $\hat{V}'(\lambda)$ since λ^* is the point at which the graph first breaks above the horizontal axis. The lower function in each graph of Figure 10 is $\hat{V}'(\lambda; \psi_{3.0})$ for these same 1WN samples. In Figure 10e $\hat{V}'(\lambda)$ bends away from zero enough so that it doesn't become positive until almost $2\lambda_0$. In Figure 10f $\hat{V}'(\lambda)$ stays below zero until λ is very much too large. A method is needed which substantially reduces λ^* in these extreme cases but has little effect on λ^* the rest of the time.

The best explanation for why $\hat{V}'(\lambda)$ sometimes lingers below zero for so long seems to involve $1/n \sum \lambda y_i \psi''(\lambda y_i)$. While each sample average in $\hat{V}(\lambda)$ converges to the corresponding expectation in $V(\lambda)$, the convergence of $1/n \sum \lambda y_i \psi''(\lambda y_i)$ appears to be the slowest. This is because the function $x \psi''(x)$, shown in Figure 11, has a relatively large total variation in comparison to the other functions in $V'(\lambda)$. If a larger number of y_i than expected are in the range where

Figure 9

Graphs of $\hat{V}(\lambda)$ Versus $\lambda \cdot MAD$ Using $\psi_{3.0}$ for Random Samples of Size 20 From the One Wild Normal

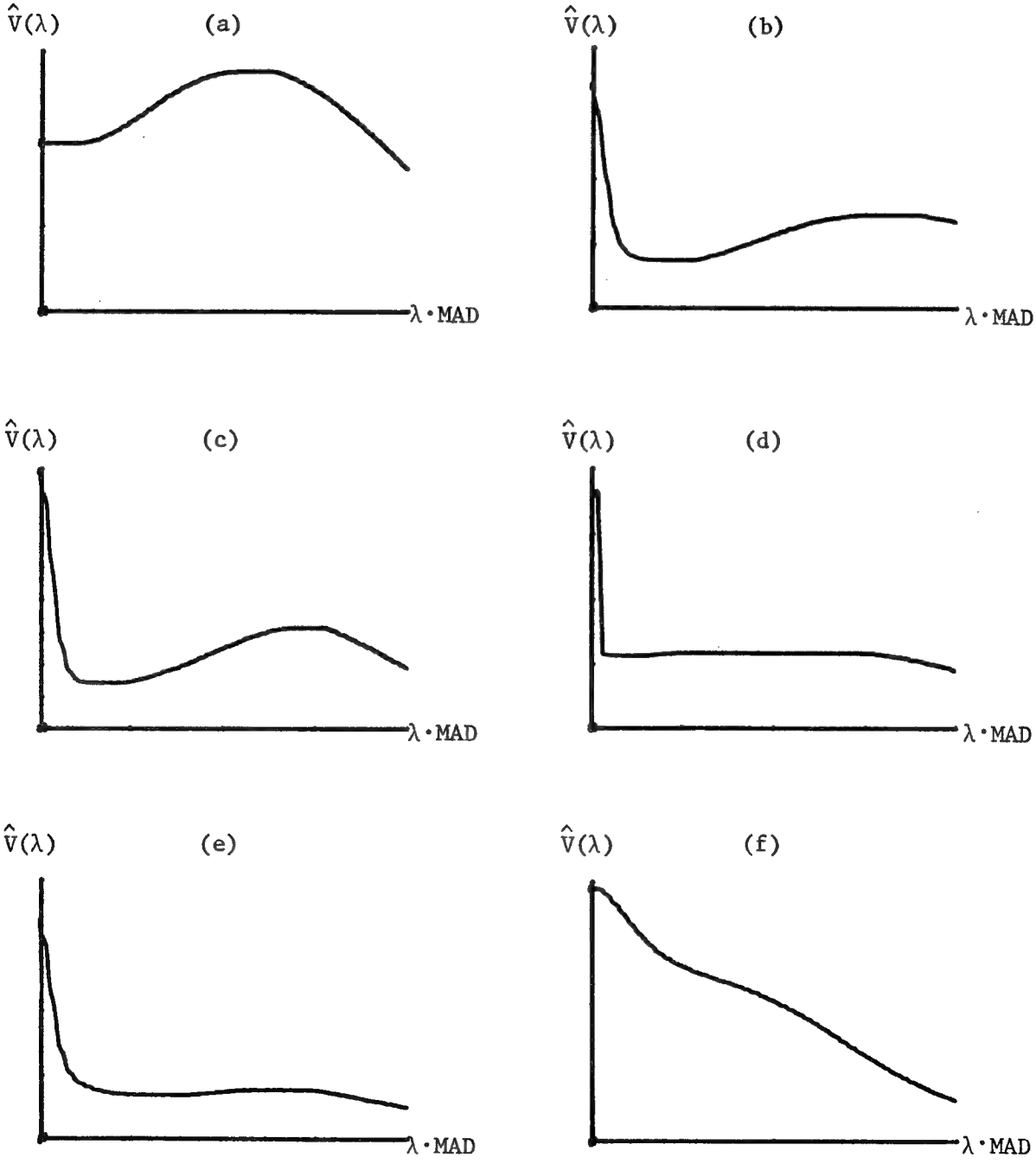


Figure 10

Graphs of $\hat{V}'(\lambda)$ and $\hat{V}'(\lambda) + g_n(\lambda)$ Versus $\lambda \cdot MAD$
Using $\psi_{3.0}$ for Random Samples of Size 20
From the One Wild Normal

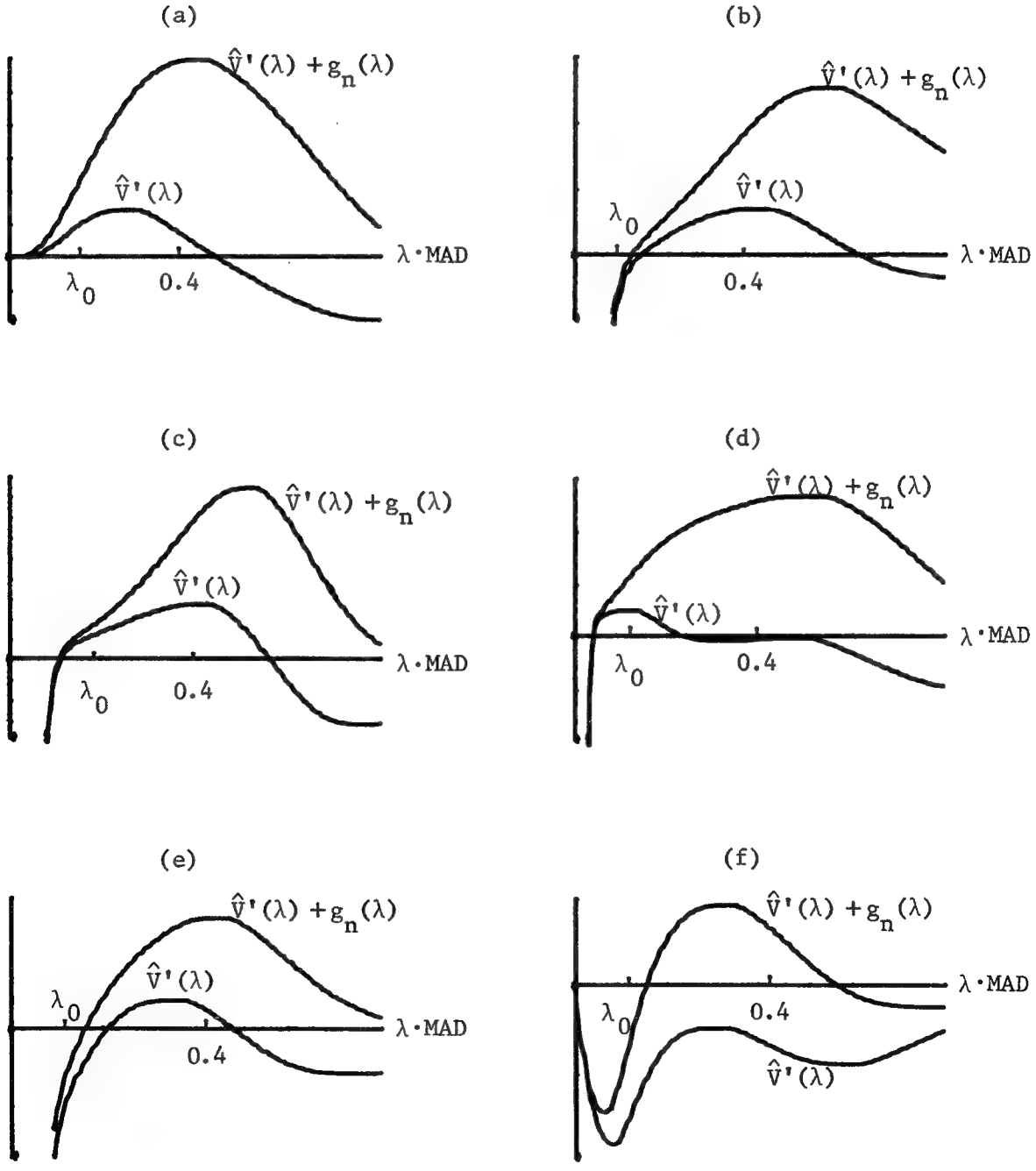
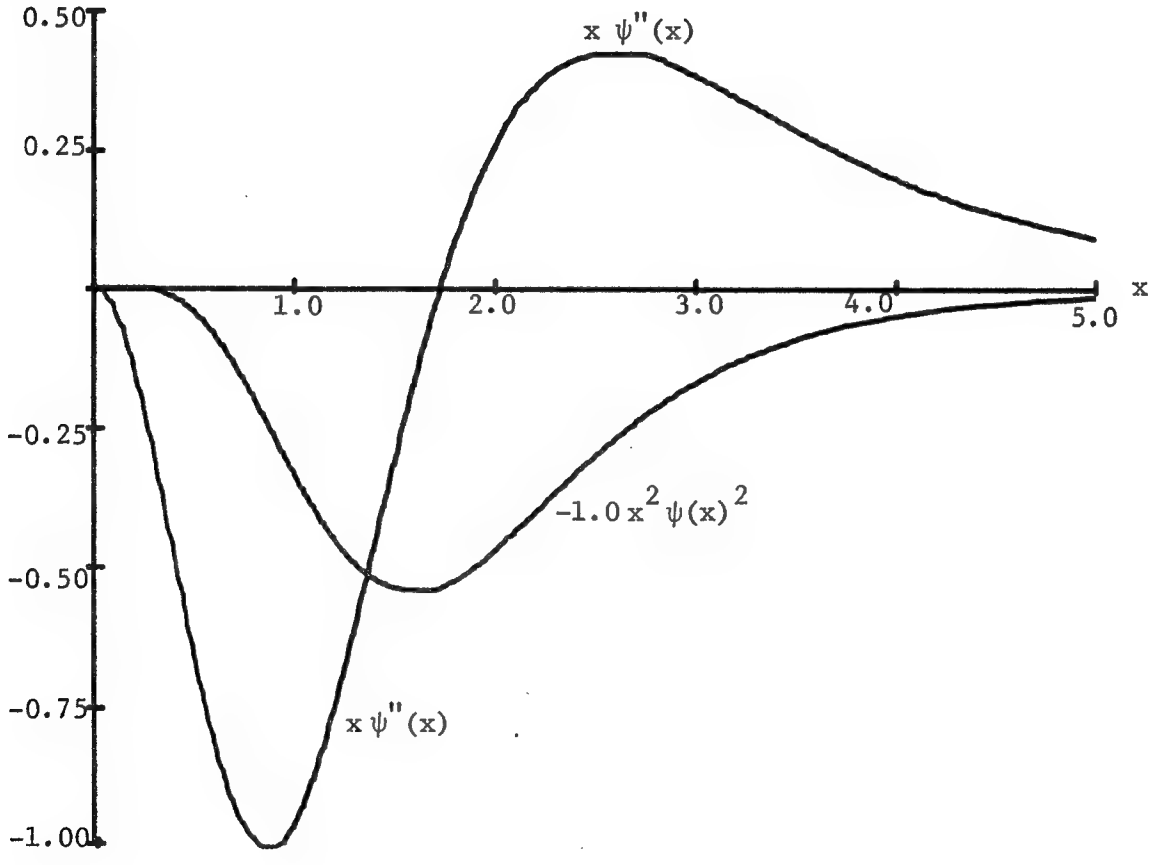


Figure 11

Graphs of $x\psi''_{3.0}(x)$ and $-1.0x^2\psi_{3.0}(x)^2$



$\lambda y\ \psi''(\lambda y_i)$ is maximized and fewer than expected where it is minimized, then $\frac{1}{n} \Sigma\ \lambda y_i\ \psi''(\lambda y_i)$ will overestimate $E\ \lambda(X-\theta)\psi''[\lambda(X-\theta)]$. In this case $\hat{V}'(\lambda)$ may remain negative well past λ_0. The estimator is improved substantially by replacing $\frac{1}{n} \Sigma\ \lambda y_i\ \psi''(\lambda y_i)$ in (7.9) with $\frac{1}{n} \Sigma[\lambda y_i\ \psi''(\lambda y_i) - c_n(\lambda y_i)^2\ \psi(\lambda y_i)^2]$, where c_n is a constant depending on the sample size n. The function $x^2\ \psi(x)^2$ is also shown in Figure 11.

We redefine λ^* by

$$\lambda^* = \inf\{\lambda > 0 : \hat{V}'(\lambda) + g_n(\lambda) > 0\} \tag{8.1}$$

where

$$g_n(\lambda) = c_n \frac{2n\ \Sigma\ \psi(\lambda y_i)^2\ \Sigma\ (\lambda y_i)^2\ \psi(\lambda y_i)^2}{\lambda^3[\Sigma\ \psi'(\lambda y_i)]^3} \ . \tag{8.2}$$

Figure 10 shows some examples of the effect of $g_n(\lambda)$ on λ^* for samples from the 1WN. In each graph the lower function is $\hat{V}'(\lambda)$ and the upper one is $\hat{V}'(\lambda) + g_n(\lambda)$ for $c_{20} = 1.0$. In the first four samples λ^* given by (7.7) is either less than λ_0 or very near it. In each of these cases, the presence of $g_n(\lambda)$ has almost no effect on λ^*. However, when λ^* given by (7.7) is much larger than λ_0, the effect of using (8.1) is substantial. In particular, it is by far the greatest in situations like that of 10f. The value of c_n can be used to tune the adaptive estimator. Increasing c_n improves the estimator for the normal and 1WN at the cost of increased variance for the slash. For sample size 20 a value of c_{20} of about 1.0 gives a reasonable balance

between short and long tailed distributions. If $c_n \to 0$ as $n \to \infty$, then the effect of g_n is asymptotically negligible. Also the presence of $g_n(\lambda)$ does not affect whether or not $\lambda^* = 0$ since $g_n(\lambda) = 0(\lambda^3)$ as $\lambda \to 0$.

One final modification has been made to the definition of λ^*. Despite utilization of the function $g_n(\lambda)$, occasionally the procedure breaks down to the extent that λ^* is greater than the reciprocal of the sample MAD. To avoid ever having more than half of the data points fall on the redescending part of the influence curve, we make the restriction $\lambda^* \leq 1/\text{MAD}$. This restriction effects less than one percent of samples of size 20.

9. Asymptotic Optimality of λ^*

In this section we show that for most common symmetric distributions, λ_n^* is asymptotically equivalent to the best scale factor λ_0. Under certain restrictions--mainly on the smoothness of ψ--Theorems 2 and 4 establish the consistency and asymptotic normality of λ_n^*, when λ_0 is positive. The asymptotic normality of $\hat{\theta}_n^*$, with best possible limiting variance using ψ, is proved in Theorem 5. Finally when F is normal, Theorems 6 and 7 establish the asymptotic efficiency of $\hat{\theta}_n^*$. First, however, we show that $\hat{\theta}^*$ and $\hat{\theta}_{(1)}^*$ are location and scale equivariant.

Theorem 1. For λ^* given by either (7.7) or (8.1) the M-estimator $\hat{\theta}^*$ and one-step M-estimator $\hat{\theta}_{(1)}^*$ are location and scale equivariant.

Proof. We only need to show that for $a > 0$, $\lambda^*(a\underset{\sim}{x}+b\underset{\sim}{1}) = \lambda^*(\underset{\sim}{x})/a$. We have immediately from (7.8) that $\hat{V}'(\lambda; a\underset{\sim}{x}+b\underset{\sim}{1}) = \hat{V}'(\lambda; a\underset{\sim}{x}) = a^3 \hat{V}'(a\lambda; \underset{\sim}{x})$. Thus the infimum of λ such that $\hat{V}'(\lambda; \underset{\sim}{x}) > 0$ is exactly a times that for $\hat{V}'(\lambda; a\underset{\sim}{x}+b\underset{\sim}{1})$, and $\lambda^*(a\underset{\sim}{x}+b\underset{\sim}{1}) = \lambda^*(\underset{\sim}{x})/a$ for λ^* defined by (7.7).

Since $g_n(\lambda; a\underset{\sim}{x}+b\underset{\sim}{1}) = a^3 g_n(a\lambda; \underset{\sim}{x})$, it follows that the theorem is also true for λ^* given by (8.1). □

In Theorem 2 we establish conditions under which λ_n^* converges almost surely to the value of λ which minimizes $V(\lambda; \psi, F)$--subject to the requirement that the first relative minimum of $V(\lambda)$ is also the absolute minimum. Fortunately, this restriction seems to be unimportant in practice. Conditions (i), (ii), and (iii) on $V'(\lambda)$ and $V''(\lambda)$ are necessary to eliminate certain pathological cases from consideration. Condition (iv), which requires a moment for the third derivative of ψ, is probably stronger than necessary. However, the class of functions satisfying (iv) apparently includes an ample choice of shapes for ψ(e.g., ψ_p given by (7.9)). Furthermore, the results in the next section suggest that a very smooth function ψ is preferable when λ is chosen adaptively. Conditions (v) and (vi) appear to be satisfied easily.

The property to follow is used in Theorems 2, 4, and 5. Let $h(\lambda,q;x)$ be a real valued function for $\lambda \in R^+$ and $q,x \in R$. For $k = 0,1,\ldots$ and $A \subset R^+$ define $h(\lambda,q;x)$ to be in the set $\mathcal{H}^k(A)$ if there exists some $\zeta > 0$ such that

$$(9.1)\qquad E \sup_{\lambda\in A,\,|q-\theta|<\zeta} \frac{\partial^j}{\partial\lambda^j} h(\lambda,q;X) < \infty\ , \qquad \text{for} \quad 0 \le j \le k+2$$

$$(9.2)\qquad E \sup_{\lambda\in A,\,|q-\theta|<\zeta} \frac{\partial^{j+1}}{\partial\lambda^j\,\partial q} h(\lambda,q;X) < \infty\ , \qquad \text{for} \quad 0 \le j \le k+1$$

and

$$(9.3)\qquad E \sup_{\lambda\in A,\,|q-\theta|<\zeta} \frac{\partial^{j+2}}{\partial\lambda^j\,\partial q^2} h(\lambda,q;X) < \infty\ , \qquad \text{for} \quad 0 \le j \le k\ .$$

Theorem 2. For symmetric F, let ψ and $\lambda_0 > 0$ be such that the conditions to follow hold. Then λ_n^* defined in (7.7) converges to λ_0 almost surely as $n \to \infty$.

(i) $V'(\lambda_0;\ \psi,F) = 0$ and $V''(\lambda_0;\ \psi,F) > 0$.

(ii) For all $\delta > 0$, $\sup_{\delta\le\lambda\le\lambda_0-\delta} V'(\lambda;\ \psi,F) < 0$.

(iii) If $E(X-\theta)^2 < \infty$, then $[E(X-\theta)^2]^2 < 1/3\ E(X-\theta)^4 \le +\infty$.

(iv) For all $\delta > 0$, $h_1(\lambda,q;\ x) = \psi[\lambda(x-q)]^2$ and $h_2(\lambda,q;\ x) = \psi'[\lambda(x-q)]$ are elements of $\mathcal{H}^0[(\delta,\infty)]$.

(v) There exists a constant $\tau > 0$, such that $\psi(x)^4$ and $[\max(x\psi''(x),\ 0)]^2$ are less than or equal to $\tau\cdot\psi(x)[\psi(x) - x\psi'(x)]$ for all x.

(vi) M converges almost surely to θ.

Proof. Since λ^* is location invariant (i.e., $\lambda^*(\underset{\sim}{x} + b\cdot\underset{\sim}{1}) = \lambda^*(x)$), we may assume that $\theta = 0$.

The proof consists of showing that for any $\varepsilon > 0$, there exists a $\delta > 0$ such that with probability one, each of the following events occurs only finitely often:

(a) $\lambda_n^* > \lambda_0 + \varepsilon$.

(b) $\delta < \lambda_n^* < \lambda_0 - \varepsilon$.

(c) $0 \leq \lambda_n^* \leq \delta$.

All subsequent statements of convergence of random variables or events refer to almost sure convergence as $n \to \infty$.

We note that for $\lambda \leq \lambda_n^*$, $\Sigma\, \psi'(\lambda y_i) > 0$. Otherwise the continuity of ψ' would imply that $\hat{V}(\ell)$ was unbounded for $\ell < \lambda$ and that $\lambda_n^* < \lambda$. Since $\Sigma\, \psi'(\lambda y_i) > 0$ for $\lambda \leq \lambda_n^*$, we can write

$$\lambda_n^* = \inf\{\lambda > 0 : W_n(\lambda) > 0\} \tag{9.4}$$

where

$$W_n(\lambda) = \frac{\left[\frac{1}{n} \Sigma\, \psi'(\lambda y_i)\right]^3 \hat{V}'(\lambda)}{2\lambda} \tag{9.5}$$

$$= \frac{1}{\lambda^4} \{\frac{1}{n} \Sigma\, \psi'(\lambda y_i) \frac{1}{n} \Sigma\, \psi(\lambda y_i) [\lambda y_i\, \psi'(\lambda y_i) - \psi(\lambda y_i)]$$

$$- \frac{1}{n} \Sigma\, \psi(\lambda y_i)^2 \frac{1}{n} \Sigma\, \lambda y_i\, \psi''(\lambda y_i)\} \quad .$$

Suppose that h is any of the functions which are summed in (9.5)--i.e., $h(t) = \psi(t)^2$, $\psi(t)[t\psi'(t) - \psi(t)]$, etc. For fixed $\lambda > 0$ and sufficiently small $\zeta > 0$, condition (iv) implies that

$E \sup_{|q|<\zeta} |\partial/\partial q\, h[\lambda(X-q)]| < \infty$. By the strong law of large numbers we have that $\overline{\lim_{n\to\infty}}\ 1/n \sum \sup_{|q|<\zeta} |\partial/\partial q\, h[\lambda(x_i - q)]| < \infty$ almost surely, which implies that $\lim_{n\to\infty} 1/n \sum \sup_{|q|<\zeta} |h[\lambda(x_i - q)] - h(\lambda x_i)| \to 0$ as $\zeta \to 0$. Since $|M_n| \to \theta = 0$ as $n \to \infty$, application of the strong law of large numbers to $1/n\ \Sigma h(\lambda x_i)$ gives us that $1/n\ \Sigma h(\lambda y_i) \to Eh(\lambda X)$. Thus for any fixed $\lambda > 0$ such that $E\ \psi'(\lambda X) > 0$, we have

$$W_n(\lambda) \to \frac{[E\ \psi'(\lambda X)]^3}{2\lambda}\ V'(\lambda) \quad . \tag{9.6}$$

(a) Let $\varepsilon > 0$ be fixed. By (i) there exists some λ, $\lambda_0 \le \lambda \le \lambda_0 + \varepsilon$, such that $V'(\lambda) > 0$ and $E\ \psi'(\lambda X) > 0$. Thus $W_n(\lambda)$ is eventually positive, and λ_n^* is eventually less than $\lambda_0 + \varepsilon$.

(b) Let $\delta > 0$ be fixed and define $\xi = \sup_{\delta \le \lambda \le \lambda_0 - \varepsilon} \frac{[E\ \psi'(\lambda X)]^3\ V'(\lambda)}{2\lambda}$. By (ii) $\xi < 0$, so that for any finite set $\{\lambda_i\}$ such that $\delta = \lambda_1 < \lambda_2 < \ldots < \lambda_k = \lambda_0 - \varepsilon$, $\sup_{1 \le i \le k} W_n(\lambda_i) < \xi/2 < 0$ for sufficiently large n. Condition (iv) and the SLLN imply there is a constant C such that $|W_n'(\lambda)| \le C$ for all $\lambda \ge \delta$. Thus if $\max_{1 \le i \le k-1} (\lambda_{i+1} - \lambda_i) < \xi/C$, then $\overline{\lim_{n\to\infty}} \sup_{1 \le i \le k} W_n(\lambda_i) < \xi/2$ implies $\overline{\lim_{n\to\infty}} \sup_{\delta < \lambda < \lambda_0 - \varepsilon} W_n(\lambda) < 0$, which implies λ_n^* is in the interval $(\delta, \lambda_0 - \varepsilon)$ only finitely often.

Note that (a) and (b) imply that there is at least one root of $V_n'(\lambda) = 0$ consistent for λ_0. We will proceed to show that there are no roots in a small neighborhood of zero.

(c) We note that $W_n(\lambda)$ may be written as

$$W_n(\lambda) = \frac{1}{n} \sum_{i=1}^{n} a(\lambda; y_i) \frac{1}{n} \sum_{i=1}^{n} b(\lambda; y_i) - \frac{1}{n} \sum_{i=1}^{n} c(\lambda; y_i) \frac{1}{n} \sum_{i=1}^{n} d(\lambda; y_i) \quad , \tag{9.7}$$

where

$$a(\lambda;y) = \psi(\lambda y)^2/\lambda^2 \quad , \tag{9.8}$$

$$b(\lambda;y) = -\lambda y\, \psi''(\lambda y)/\lambda^2 \quad , \tag{9.9}$$

$$c(\lambda;y) = \psi'(\lambda y) \quad , \tag{9.10}$$

and

$$d(\lambda;y) = \frac{\psi(\lambda y)}{\lambda} \frac{[\psi(\lambda y) - \lambda y\, \psi'(\lambda y)]}{\lambda^3} \quad . \tag{9.11}$$

We need to show that there exists some $\delta > 0$ such that for n sufficiently large, $\sup_{0<\lambda\leq\delta} W_n(\lambda) < 0$.

Let $\rho > 0$ be arbitrarily small. Since $\psi'(0) = 1$ and ψ' is continuous at zero and bounded below, there exists some $\delta > 0$ such that $\lim_{n\to\infty} \frac{1}{n} \Sigma_{i=1}^{n} c(\lambda; y_i) \geq 1 - \rho$ uniformly for $\lambda \leq \delta$. Thus since (v) implies $d(\lambda;y) \geq 0$, we have that

$$W_n(\lambda) \leq \frac{1}{n} \sum_{i=1}^{n} a(\lambda;y_i) \frac{1}{n} \sum_{i=1}^{n} b(\lambda;y_i) - (1-\rho) \frac{1}{n} \Sigma\, d(\lambda;y_i) \tag{9.12}$$

for $\lambda \leq \delta$ and n sufficiently large.

At this point we separate each of the summations in (9.12) into two parts according to whether $y_i \leq B$ (for a large positive number B). Along with the condition on δ from the previous paragraph, we require that $\delta \leq \eta/B$ where $\eta > 0$ and B are constants to be determined later. Thus for sufficiently large n

$$(9.13)\quad W_n(\lambda) \le \left[\frac{1}{n}\sum_{y_i\le B} a(\lambda;y_i) + \frac{1}{n}\sum_{y_i>B} a(\lambda;y_i)\right]$$

$$\cdot \left[\frac{1}{n}\sum_{y_i\le B} b(\lambda;y_i) + \frac{1}{n}\sum_{y_i>B} b(\lambda;y_i)\right]$$

$$- (1-\rho)\left[\frac{1}{n}\sum_{y_i\le B} d(\lambda;y_i) + \frac{1}{n}\sum_{y_i>B} d(\lambda;y_i)\right] .$$

Since $y_i \le B$ implies that $\lambda y_i \le \delta B \le \eta$, Taylor expansions of the terms in (9.13) for $y_i \le B$ give

$$(9.14)\quad W_n(\lambda) \le \left([1+0(\eta^2)]\frac{1}{n}\sum_{y_i\le B} y_i^2 + \frac{1}{n}\sum_{y_i>B} a(\lambda;y_i)\right)$$

$$\cdot \left([-\psi'''(0)+0(\eta^2)]\frac{1}{n}\sum_{y_i\le B} y_i^2 + \frac{1}{n}\sum_{y_i>B} b(\lambda;y_i)\right)$$

$$- (1-\rho)\left([-\tfrac{1}{3}\psi'''(0)+0(\eta^2)]\frac{1}{n}\sum_{y_i\le B} y_i^4 + \frac{1}{n}\sum_{y_i>B} d(\lambda;y_i)\right)$$

for sufficiently large n. Remember that $\psi'''(0) < 0$.

Finally we must consider separately the cases $EX^2 < \infty$ and $EX^2 = +\infty$. First suppose $EX^2 < \infty$. Then for sufficiently large n, we have

$$(9.15)\quad \sup_{0<\lambda\le\delta} W_n(\lambda) \le \left([1+0(\eta^2)]\frac{1}{n}\sum_{y_i\le B} y_i^2 + \frac{1}{n}\sum_{y_i>B} y_i^2\right)$$

$$\cdot [-\psi'''(0)]\left([1+0(\eta^2)]\frac{1}{n}\sum_{y_i\le B} y_i^2 + \frac{1}{n}\sum_{y_i>B} y_i^2\right)$$

$$- (1-\rho)[-\tfrac{1}{3}\psi'''(0)+0(\eta^2)]\frac{1}{n}\sum_{y_i\le B} y_i^4 .$$

For fixed B, the same reasoning which leads to (9.6) implies that the summations $1/n \Sigma_{y_i \leq B} y_i^2$, $1/n \Sigma_{y_i > B} y_i^2$, and $1/n \Sigma_{y_i \leq B} y_i^4$ converge to $\int_{|x| \leq B} x^2 \, dF(x)$, $\int_{|x|>B} x^2 \, dF(x)$, and $\int_{|x| \leq B} x^4 \, dF(x)$ respectively as $n \to \infty$. As $B \to \infty$, the three integrals converge to EX^2, 0, and EX^4. Finally since $1/3 \, EX^4 > (EX^2)^2$, choosing ρ and η sufficiently small, and B and n sufficiently large, implies that $\sup_{0<\lambda \leq \delta} W_n(\lambda) < 0$.

Now suppose that $EX^2 = +\infty$. By reasoning similar to that above we can choose ρ, η, and B_0 in such a way that if $B \geq B_0$, then there exists some $\gamma > 0$ for which $\Sigma_{y_i>B} a(\lambda;y_i) \leq \gamma \Sigma_{y_i \leq B} y_i^2$ and $\Sigma_{y_i>B} b(\lambda;y_i) \leq \gamma \Sigma_{y_i \leq B} y_i^2$ implies $\sup_{0<\lambda \leq \delta} W_n(\lambda)$ is eventually negative. Thus we only need to worry about those n and $\lambda < \delta$ such that one of the inequalities fails to hold.

Suppose, for example, that $\Sigma_{y_i>B} a(\lambda;y_i) > \gamma \Sigma_{y_i \leq B} y_i$. Let the random variable p_n equal the number of $y_i > B$. Then (v) implies that $1/n \Sigma_{y_i>B} d(\lambda;y_i) \geq (p_n/\tau_n) \; 1/p_n \Sigma_{y_i>B} a(\lambda;y_i)^2 \geq p_n/\tau_n [1/p_n \Sigma_{y_i>B} a(\lambda;y_i)]^2 = n/\tau p_n [1/n \Sigma_{y_i>B} a(\lambda;y_i)]^2$. It is possible to choose B large enough so that $P(|X_i| > B) \leq \gamma/6\tau$. In that case p_n/n is eventually less than $\gamma/3\tau$ which implies that

$$\frac{1}{3n} \sum_{y_i>B} d(\lambda;y_i) \geq \frac{1}{\gamma} \left[\frac{1}{n} \sum_{y_i>B} a(\lambda;y_i) \right]^2 \geq \frac{1}{n} \sum_{y_i \leq B} y_i^2 \, \frac{1}{n} \sum_{y_i>B} a(\lambda;y_i) \; .$$

Similarly we get that $\frac{1}{3n} \Sigma_{y_i>B} d(\lambda;y_i) \geq \frac{1}{n} \Sigma_{y_i \leq B} y_i^2 \, \frac{1}{n} \Sigma_{y_i>B} b(\lambda;y_i)$ and $\frac{1}{3n} \Sigma_{y_i>B} d(\lambda;y_i) \geq \frac{1}{\gamma} \frac{1}{n} \Sigma_{y_i>B} a(\lambda;y_i) \, \frac{1}{n} \Sigma_{y_i>B} b(\lambda;y_i)$. Since

these inequalities hold for all $\lambda \in (0,\delta)$ (for large enough n), we have $\sup_{0<\lambda\leq\delta} W_n(\lambda) \geq 0$ only finitely often and thus $\lambda_n^* \leq \delta$ only finitely often. □

The desired consequence of Theorem 2 is that $\sqrt{n}(\hat{\theta}_n^* - \theta)$ has a limiting normal distribution with variance $V(\lambda_0; \psi, F)$, the best possible asymptotic variance using ψ. Before proving that result, as Theorem 5, we need to study the rate of convergence of λ_n^* to λ_0. In Theorem 4 we show, under slightly more restrictive conditions than in Theorem 2, that $\sqrt{n}(\lambda_n^* - \lambda_0)$ has a limiting normal distribution. With the use of this intermediate result, we are able to obtain the asymptotic optimality of λ_n^* as a scale factor when $\lambda_0 > 0$. The most interesting case for which $\lambda_0 = 0$ occurs when F is the normal distribution function. Theorems 6 and 7 give the asymptotic optimality of λ_n^* and θ_n^* when F is normal.

Theorems 4-7 share several common elements. In each theorem we study the limiting behavior of the root of an implicit equation in the presence of a nuisance parameter which converges to a constant in probability. In Theorems 5 and 7, where the implicit equation is the defining equation for an M-estimate, the results are rather familiar. Because the defining equation for λ_n^* is more complex, Theorems 4 and 6 require a more general framework. Theorem 3 is a central limit theorem in this more general setup. No attempt has been made to find the most general conditions under which the theorem is true. As a consequence, Theorems 4-7 may call for more derivatives of ψ than are absolutely necessary. However, for the same reasons that precede the

statement of Theorem 2, this limitation should have little significance.

Theorem 3 requires the following definitions. As before, suppose that $x_1,\ldots,x_n$ is a random sample from F, and let q_n be a real valued function of $\underset{\sim}{x}$ such that $q_n(\underset{\sim}{x})$ converges in probability to the constant q_0. Given the real valued functions $a_j(t,q;x)$, for $j = 1,\ldots,4$; let

$$U_n(t) = \frac{1}{n}\sum_{i=1}^{n} a_1(t,q_n;x_i)\,\frac{1}{n}\sum_{i=1}^{n} a_2(t,q_n;x_i) + \frac{1}{n}\sum_{i=1}^{n} a_3(t,q_n;x_i)\,\frac{1}{n}\sum_{i=1}^{n} a_4(t,q_n;x_i) \quad . \tag{9.16}$$

and

$$U(t) = Ea_1(t,q_0;X)Ea_2(t,q_0;X) + Ea_3(t,q_0;X)Ea_4(t,q_0;X) \quad . \tag{9.17}$$

Lemma 1 gives the limiting distribution of T_n, a consistent root of the equation,

$$U_n(T_n) = 0 \quad . \tag{9.18}$$

Finally, define

$$\sigma^2 = \underset{\sim}{\alpha}^t D \underset{\sim}{\alpha} \tag{9.19}$$

where

$$\underset{\sim}{\alpha} = (Ea_2(t_0,q_0;X), Ea_1(t_0,q_0;X), Ea_4(t_0,q_0;X), Ea_3(t_0,q_0;X))^t \tag{9.20}$$

and D is the covariance matrix of $(a_1(t_0,q_0;X), a_2(t_0,q_0;X), a_3(t_0,q_0;X), a_4(t_0,q_0;X))$.

Theorem 3. Under the conditions C1-C6, there exists a consistent sequence T_n of eventually unique roots of (9.18) converging to t_0 as $n \to \infty$, and $\sqrt{n}(T_n - t_0) \xrightarrow{\mathcal{D}} N(0,\sigma^2/[U'(t_0)]^2)$.

(C1) $U(t_0) = 0$.

(C2) $U'(t_0) \neq 0$.

(C3) $q_n = q_0 + O_p(n^{-\frac{1}{2}})$.

(C4) $E\ \partial/\partial q\ a_j(t_0,q_0;X) = 0$, for $j = 1,\ldots,4$.

(C5) There exist neighborhoods T of t_0 and Q of q_0 such that $E \sup_{T,Q} |\partial^2/\partial t^2\ a_j(t,q;X)|$, $E \sup_{T,Q} |\partial^2/\partial t \partial q\ a_j(t,q,X)|$, and $E \sup_{T,Q} |\partial^2/\partial q^2\ a_j(t,q;X)|$ are finite for $j = 1,\ldots,4$.

(C6) The covariance matrix D is finite.

Proof. The first step is to show that $\sqrt{n}\ U_n(t_0)$ has a limiting normal distribution. Expanding $a_j(t_0,q_n;x_i)$ in a Taylor series about $q_n = q_0$ gives

$$(9.21)\quad \frac{1}{n}\sum_{i=1}^{n} a_j(t_0,q_n;x_i) = \frac{1}{n}\Sigma\, a_j(t_0,q_0;x_i) + (q_n-q_0)\frac{1}{n}\Sigma\frac{\partial}{\partial q}a_j(t_0,q_0;x_i) + \frac{1}{2}(q_n-q_0)^2\frac{1}{n}\Sigma\frac{\partial^2}{\partial q^2}a_j(t_0,r_n;x_i)$$

for some r_n such that $|r_n - q_0| \le |q_n - q_0|$. Since

$$\sup_{|r_n-q_0|\le|q_n-q_0|} |1/n \Sigma\, \partial^2/\partial q^2\, a_j(t_0,r_n;x_i)|$$

$$\le 1/n \Sigma \sup_{|r_n-q_0|\le|q_n-q_0|} |\partial^2/\partial q^2\, a_j(t_0,r_n;x_i)|,$$ (C3), (C5), and the law of large numbers imply that $1/n \Sigma\, \partial^2/\partial q^2\, a_j(t_0,r_n;x_i) = O_p(1)$ as $n \to \infty$. Using C3 again implies that the last term of (9.21) is $O_p(n^{-1}) = o_p(n^{-\frac{1}{2}})$. By C3, C5, and the law of large numbers the second term is also $o_p(n^{-\frac{1}{2}})$. Thus the multivariate central limit theorem implies that

$$(9.22) \qquad \sqrt{n} \begin{pmatrix} \frac{1}{n} \Sigma\, a_1(t_0,q_n;x_i) - E\, a_1(t_0,q_0;X) \\ \vdots \\ \frac{1}{n} \Sigma\, a_4(t_0,q_n;x_i) - E\, a_4(t_0,q_0;X) \end{pmatrix} \xrightarrow{\mathcal{D}} N_4(\underset{\sim}{0},D) \quad .$$

Condition C1 now implies that

$$(9.23) \qquad \sqrt{n}\, U_n(t_0) \xrightarrow{\mathcal{D}} N(0,\sigma^2) \quad .$$

We next consider the behavior of $U_n'(t)$ in T, the neighborhood of t_0 mentioned in C5. We can write

$$(9.24) \qquad \frac{1}{n} \Sigma \frac{\partial}{\partial t} a_j(t,q_n;x_i) = \frac{1}{n} \Sigma \frac{\partial}{\partial t} a_j(t,q_0;x_i)$$

$$+ (q_n - q_0) \frac{1}{n} \Sigma \frac{\partial^2}{\partial t \partial q} a_j(t,r_n;x_i)$$

for some new value of r_n such that $|r_n - q_0| \le |q_n - q_0|$. For $t \in T$, the second term on the right is $O_p(n^{-\frac{1}{2}})$, so that a further Taylor expansion about $t = t_0$ gives

$$(9.25)\quad \frac{1}{n}\sum \frac{\partial}{\partial t} a_j(t,q_n;x_i) = \frac{1}{n}\sum \frac{\partial}{\partial t} a_j(t_0,q_0;x_i) + (t-t_0)\sum \frac{\partial^2}{\partial t^2} a_j(v_n,q_0;x_i) + O_p(n^{-\frac{1}{2}})$$

for some $v_n \in T$. Near repetition of the steps following (9.21) implies that

$$(9.26)\quad \frac{1}{n}\sum \frac{\partial}{\partial t} a_j(t,q_n;x_i) = E\frac{\partial}{\partial t} a_j(t_0,q_0;X) + O(|t-t_0|) + o_p(1)\ ,$$

and consequently

$$(9.27)\qquad U_n'(t) = U(t_0) + O(|t-t_0|) + o_p(1)\ ,$$

where $O(|t-t_0|)$ does not depend on n.

Thus for any $\varepsilon > 0$, there exists some $\delta > 0$ such that $|t-t_0| < \delta$ implies that

$$(9.28)\qquad P(|U_n'(t) - U'(t_0)| > \varepsilon) \to 0$$

as $n \to \infty$. By choosing $\varepsilon < |U'(t_0)|$, we can find an interval around t_0 such that the probability of multiple roots converges to zero. Furthermore, (9.23) insures that the probability of there being a root in the interval converges to one. Finally, (9.28) implies that for any $\varepsilon > 0$

$$(9.29)\qquad P\left[1 - \frac{\varepsilon}{|U'(t_0)|} \le -\frac{U_n(t_0)}{U'(t_0)(T_n - t_0)} \le 1 + \frac{\varepsilon}{|U'(t_0)|}\right] \to 0$$

as $n \to \infty$. Thus $\sqrt{n}\,(T_n - t_0)$ must have the same limiting distribution as $-\sqrt{n}\,U_n(t_0)/U'(t_0)$. □

Theorem 3 reduces the proofs of Theorems 4-7 to little more than condition checking. We note the complementary relationship between t and q. In each theorem the random variable of primary interest is t. In Theorems 4 and 6, when t is the scale factor (or its square), q is a location parameter--the median. In contrast, when t is a location estimate, q is a function of the scale factor.

We define several functions and constants in preparation for Theorem 5. Let

$$Q_1(x) = \psi[\lambda_0(x-\theta)]^2 \tag{9.30}$$

$$Q_2(x) = -\lambda_0(x-\theta)\psi''[\lambda_0(x-\theta)] \tag{9.31}$$

$$Q_3(x) = \psi'[\lambda_0(x-\theta)], \text{ and} \tag{9.32}$$

$$Q_4(x) = \psi[\lambda_0(x-\theta)]\{\lambda_0(x-\theta)\psi''[\lambda_0(x-\theta)] - \psi[\lambda_0(x-\theta)]\} . \tag{9.33}$$

Then define

$$\sigma^2 = \underset{\sim}{\alpha}^t D \underset{\sim}{\alpha} \tag{9.34}$$

where

$$\underset{\sim}{\alpha} = (EQ_2(X), EQ_1(X), EQ_4(X), EQ_3(X))^t \tag{9.35}$$

and D is the covariance matrix of $\underset{\sim}{Q}(X)$. Also define

$$\omega = \frac{1}{2}\lambda_0^3\{E\psi'[\lambda_0(X-\theta)]\}^3 V''(\lambda_0) \quad . \tag{9.36}$$

Theorem 4. Suppose that ψ, F, and λ_0 satisfy the conditions below in addition to those of Theorem 2. Then for λ_n^* given by (7.7)

(9.37) $$\sqrt{n}\ (\lambda_n^* - \lambda_0) \xrightarrow{\mathcal{D}} N(0, \frac{\sigma^2}{\omega^2}) \quad .$$

(iv') For some $\delta > 0$, $h_1(\lambda,q;x) = \psi[\lambda(x-q)]^2$, $h_2(\lambda,q;x) = \psi'[\lambda(x-q)]$, and $h_3(\lambda,q;x) = \psi[\lambda(x-q)]$ are elements of $\mathcal{H}^1[(\lambda_0 - \delta, \lambda_0 + \delta)]$.

(vi') $M_n = \theta + O_p(n^{-\frac{1}{2}})$.

(vii) $EQ_j(X)^2 < \infty$, for $j = 1,\ldots,4$.

Proof. Let $T_n = \lambda_n^*$, $t_0 = \lambda_0$, $q_n = M_n$, $q_0 = \theta$, $a_1(t,q;x) = \psi[t(x-q)]^2$, $a_2(t,q;x) = -t(x-q)\psi''[t(x-q)]$, $a_3(t,q;x) = \psi'[t(x-q)]$, and $a_4(t,q;x) = \psi[t(x-q)]\{t(x-q)\psi'[t(x-q)] - \psi[t(x-q)]\}$. We note that $U_n(t) = 1/2\ t^3\{1/n\ \Sigma\ \psi'[t(x_i - M_n)]\}^3\ \hat{V}'(t)$, so that indeed $U_n(T_n) = 0$. Also $U'(t_0) = 1/2\ t_0^3\{E\psi'[t_0(X-\theta)]\}^3\ V''(t_0)$, so that $U'(\lambda_0) = \omega$. Thus if conditions C1-C6 hold, application of Theorem 3 gives the desired result.

C1 and C2 each follow from condition (i) of Theorem 2. C3 is implied by assumption (vi'). C4 follows from the fact that, for each j, $\partial/\partial q\ a_j(t_0,q_0;X)$ has a symmetric distribution centered at zero. Finally, C5 and C6 are both assumed in the hypothesis of the theorem. □

Theorem 5. Under the assumptions of Theorem 4, there exists a consistent sequence of roots $\hat{\theta}_n^*$ with limiting distribution

(9.38) $$\sqrt{n}\ (\hat{\theta}_n^* - \theta) \xrightarrow{\mathcal{D}} N(0, V(\lambda_0;\psi,F)) \quad .$$

Proof. The proof follows the same line as the last one with $T_n = \hat{\theta}_n^*$, $t_0 = \theta$, $q_n = \lambda_n^*$, $q_0 = \lambda_0$, $a_1 = \psi[q(x-t)]$, $a_2 \equiv 1$, and $a_3 \equiv a_4 \equiv 0$. Conditions C1-C6 are all verified easily.

Theorem 6. Let F be a normal distribution function. If ψ has four bounded derivatives on the real line and $M_n = O_p(n^{-\frac{1}{2}})$, then $\lambda_n^* = O_p(n^{-\frac{1}{4}})$.

Proof. The proof relies on Theorem 3 and some reasoning specific to this case. Let $t_0 = 0$, $q_n = M_n$, and $q_0 = \theta$. For $t > 0$, let $a_1 = \psi[t^{\frac{1}{2}}(x-q)]/t$, $a_2 = -t^{\frac{1}{2}}(x-q)\psi''[t^{\frac{1}{2}}(x-q)]/t$, $a_3 = \psi'[t^{\frac{1}{2}}(x-q)]$, and $a_4 = \psi[t^{\frac{1}{2}}(x-q)]\{t^{\frac{1}{2}}(x-q)\psi'[t^{\frac{1}{2}}(x-q)] - \psi[t^{\frac{1}{2}}(x-q)]\}/t^2$. For $t = 0$, let $a_1(0,q;x) = (x-q)^2$, $a_2(0,q;x) = -\psi'''(0)(x-q)^2$, $a_3(0,q;x) = 1$, and $a_4(0,q;x) = 1/3\ \psi'''(0)(x-q)^4$. We would like to set $T_n = (\lambda_n^*)^2$, but it is not always true that $U_n[(\lambda_n^*)^2] = 0$. That is true only when $U_n(0) \leq 0$. However, when $U_n(0) > 0$, λ_n^* is zero and the conclusion of Theorem 6 is trivial. Thus we only need to worry about the times that $T_n = (\lambda_n^*)^2$. If conditions C1-C6 hold, then the theorem is true.

Expanding U(t) in a Taylor series about t = 0, yields

$$U(t) = -\psi'''(0)\{[E(X-\theta)^2]^2 - \tfrac{1}{3}E(X-\theta)^4\} + t\{\tfrac{1}{18}[\psi'''(0)]^2[E(X-\theta)^6 - 3E(X-\theta)^2E(X-\theta)^4] + \tfrac{1}{3}\psi^{(5)}(0)[E(X-\theta)^6 - 5E(X-\theta)E(X-\theta)^4]\} + O(t^2) \quad . \tag{9.39}$$

Since the kurtosis of the normal is zero, $U(0) = 0$ which implies C1. Because the consistent root must be positive when $U_n(0) < 0$, it is necessary that $U'(0) > 0$ (rather than just $\neq 0$). For the normal $U'(0) = 1/3[\psi'''(0)]^2[\mathrm{Var}(X)]^3 > 0$, so that C2 holds. The rest of the conditions are verified easily by using the hypotheses of the theorem. We note that the slightly weaker condition on the derivatives of ψ is made possible by the fact that all moments of the normal distribution are finite. □

Theorem 7. Under the conditions of Theorem 6,

$$\sqrt{n}\,(\hat{\theta}_n^* - \theta) \xrightarrow{\mathcal{D}} N(0, \mathrm{Var}(x)) \quad . \tag{9.40}$$

Proof. We let $T_n = \hat{\theta}_n^*$, $t_0 = \theta$, $q_n = (\lambda_n^*)^2$, $q_0 = 0$, $a_1 = \psi[q^{\frac{1}{2}}(x - t)]$, $a_2 \equiv 1$, and $a_3 \equiv a_4 \equiv 0$. Again, C1-C6 are verified easily. □

10. Monte Carlo Results

In this section we present results demonstrating the performance of the adaptive one-step M-estimator on the triefficiency sampling situations described in Section 2. Numerous preliminary Monte Carlo runs were done in the developmental stages to determine the exact form of $\hat{\theta}^*_{(1)}$ and appropriate parameter values. The basic estimator and the variations which are presented here were decided upon before the final Monte Carlo study was done. Details of the design of the study may be found in Appendix B.

Table 1 contains the main results of this section. The two main estimators being compared are:

1. basic adaptive one-step M-estimator, $\hat{\theta}^*_{(1)}$, given by (5.4) where $\lambda = \lambda^*$ is defined by (8.1). The influence curve is $\psi_{3.0}(x)$, and $c_{20} = 1.0$. Other functions ψ and values of c_{20} are used in Tables 2 and 3 respectively.

2. nonadaptive one-step bisquare, given by (5.4) where $\lambda = 1/(6.4 \cdot \text{MAD})$. This estimator (for constants 7.4, 8.2, and 9.0) is used in a Monte Carlo study of confidence interval robustness by Gross (1976). The constant 6.4 has been chosen to produce fairly equal performance (in terms of relative efficiency to the best known estimators) for the normal and slash. The efficiency of the bisquare on the 1WN is especially good. This estimator is also used in the triefficiency study of P-estimators by Johns (1979). We shall see that the particular choice of the constant is not especially important since the value of c_{20} serves to tune $\hat{\theta}^*_{(1)}$ in practically the same manner.

TABLE 1

TRIEFFICIENCY OF BASIC ADAPTIVE ONE-STEP M-ESTIMATOR FOR $p = 3.0$, $c_{20} = 1.0$, $n = 20$

	Normal	1WN	Slash
Number of samples	10,000	20,000	100,000
Variance of $\hat{\theta}^*_{(1)} \times 20$	1.070 (0.003)	1.197 (0.003)	6.172 (0.025)
Relative efficiency to "best known"	93.5 (0.3)	94.2	92.7
Relative efficiency to bisquare	105.0 (0.20)	98.9 (0.14)	103.5 (0.17)
Relative efficiency to nonadaptive one-step; $\psi_{3.0}$, $\lambda = 0.35/\text{MAD}$	103.1 (0.20)	98.3 (0.14)	101.8 (0.17)

The second line of Table 1 lists $n = 20$ times the variance of the basic adaptive estimator. Standard errors of the estimated variances appear in parentheses. In the next line those variances are compared with the best known variances of location and scale equivariant estimators. For the normal the best estimator is the mean, with variance 1. For the other two situations, the true best estimator is very difficult to calculate. For the 1WN we have used the variance 1.127, reported in the Princeton study for a Hampel, 25A. For the slash, the best variance which we could find is 5.72 for the best one-step bisquare in Figure 3. The relative efficiency of $\hat{\theta}^*_{(1)}$ is

in the neighborhood of 93-94 percent for each of the three diverse situations.

A more realistic test of a new estimator is comparison against a single estimator with known good properties. Relative efficiencies of $\hat{\theta}^*_{(1)}$ to the bisquare are given in line four of Table 1. These show that $\hat{\theta}^*_{(1)}$ substantially outperforms the bisquare on the normal (105.0) and slash (103.5) while sacrificing little on the 1WN (98.9). This is typical of the performance of the adaptive estimator. It does best for extreme distributions, where the loss from compromising is the greatest. For moderate tailed distributions, where there is little to be gained by adjusting λ, the adaptive esimator does not do as well.

In the last line of Table 1 $\hat{\theta}^*_{(1)}$ is compared to a nonadaptive one-step with the same function $\psi_{3.0}$, instead of ψ_{bs}. While the adaptive estimator still seems to offer an improvement, the gain is not nearly so striking. This is because $\psi_{3.0}$ uniformly dominates the bisquare for these three situations. This fact strongly suggests that for nonadaptive M-estimators an influence curve which asymptotically returns to zero is preferable to one which is zero for all but a finite interval of x values.

Tables 2, 3, and 4 give Monte Carlo results for modifications of the basic adaptive estimator. This allows us to examine how certain parameters affect $\hat{\theta}^*_{(1)}$. Each of these results is based on a 20 percent subset of the samples used to create Table 1. By using a carefully chosen subset of samples, we have reduced the standard

errors of most of the entries well below what they otherwise would be. Furthermore, the standard error of the difference between two entries in the same column is usually smaller than either standard error. Thus patterns appearing in the columns of the tables should closely reflect the truth. For more details on these points, see Appendix B.

In Table 2 we study the effect on the triefficiency of $\hat{\theta}^*_{(1)}$ of using various shaped functions ψ. We use the family ψ_p in (7.9) and (7.10) for p = 1.5, 2.0, 3.0, and ∞. Remember that as p increases, $\psi_p(x)$ redescends more quickly from its maximum at x = 1.0. Although the value of p has some effect on the performance of $\hat{\theta}^*_{(1)}$ (at least for the normal and slash), the adaptive estimator is not overly sensitive to changes of p in this range. However, the

TABLE 2

RELATIVE EFFICIENCY OF $\hat{\theta}^*_{(1)}$ TO THE BISQUARE FOR VARIOUS FUNCTIONS ψ; $c_{20} = 1.0$

	Normal	1WN	Slash
ψ_p , p = 1.5	105.7 (0.3)	98.5 (0.3)	102.6 (0.2)
ψ_p , p = 2.0	105.4 (0.2)	98.8 (0.2)	103.4 (0.2)
ψ_p , p = 3.0	105.0 (0.2)	98.9 (0.14)	103.5 (0.17)
ψ_p , p = ∞	104.4 (0.4)	98.8 (0.2)	102.2 (0.3)
ψ_{bs}	104.7 (0.5)	100.0 (0.4)	87.9 (0.5)

moderate values--2.0 and 3.0--do appear to be slightly superior to either $p = 1.5$ or $p = \infty$.

The last line of Table 2 shows the performance of $\hat{\theta}^*_{(1)}$ using ψ_{bs} (rescaled to have a maximum at 1.0). The relative efficiencies to the bisquare are good for the normal and 1WN but disasterous for the slash. The apparent reason for this problem is that ψ''_{bs} is discontinuous at $\pm \sqrt{5}$ (for this scaling). This topic was discussed in Section 7.

Table 3 demonstrates the effect for $p = 3.0$ and $n = 20$ of c_n, which appears in formula (8.2) for $g_n(\lambda)$. The relative efficiencies in the first line (where $c_{20} = 0.0$) correspond to defining λ^* without the function $g_n(\lambda)$. As we stated in Section 8, the results are

TABLE 3

RELATIVE EFFICIENCY OF $\hat{\theta}^*_{(1)}$ TO THE BISQUARE
FOR VARIOUS VALUES OF c_{20}; $p = 3.0$

	Normal	1WN	Slash
$c_{20} = 0.0$	94.8 (1.0)	88.1 (0.7)	109.4 (0.4)
$c_{20} = 0.9$	104.4 (0.2)	98.5 (0.15)	104.3 (0.2)
$c_{20} = 1.0$	105.0 (0.2)	98.9 (0.14)	103.5 (0.17)
$c_{20} = 1.1$	105.6 (0.3)	99.6 (0.2)	102.7 (0.2)
$c_{20} = 1.3$	106.6 (0.3)	100.5 (0.3)	100.8 (0.2)

unacceptable for the normal and 1WN. Values of c_{20} in the interval (0.9, 1.3) all appear to be reasonable choices. As c_{20} increases in this range, the efficiency of $\hat{\theta}^*_{(1)}$ improves for the normal and 1WN. At the same time the efficiency for the slash decreases. We see that c_n can be used to tune $\hat{\theta}^*_{(1)}$ to allow for prior assessment of the likelihood of very long tails. There is an essential relationship between c_n and the scaling of ψ. Different members of the family $\psi_p(x)$ gave similar results for a single value of c_{20} because each of these functions has its maximum at $x = 1.0$. If ψ_p is rescaled to have a maximum at t, then c_{20} should be divided by t^4.

In Table 4 we present results of an ad hoc modification which uniformly improves $\hat{\theta}^*_{(1)}$ for the three sampling situations. We noted in Section 7 that $\hat{V}(\lambda)$ is least reliable as an estimate of $V(\lambda)$ when $1/n \Sigma \psi'(\lambda y_i)$ is small. The modification places a lower bound on $1/n \Sigma \psi'(\lambda y_i)$. If this bound is reached before $\hat{V}'(\lambda)$ becomes

TABLE 4

RELATIVE EFFICIENCY OF $\hat{\theta}^*_{(1)}$ TO THE BISQUARE FOR VARIOUS LOWER BOUNDS ON $\frac{1}{n} \Sigma \psi'(\lambda y_i)$; $p = 3.0$, $c_{20} = 1.0$

	Normal	1WN	Slash
$\frac{1}{n} \Sigma \psi'(\lambda y_i) \geq 0.0$	105.0 (0.2)	98.9 (0.14)	103.5 (0.17)
$\frac{1}{n} \Sigma \psi'(\lambda y_i) \geq 0.40$	105.2 (0.2)	99.4 (0.2)	103.7 (0.2)
$\frac{1}{n} \Sigma \psi'(\lambda y_i) \geq 0.45$	105.6 (0.3)	100.0 (0.2)	103.1 (0.2)

positive, λ^* is set equal to the point at which this happens. For a lower bound of 0.40, small improvements of 0.2 to 0.5 percent are registered for each of the triefficiency situations. Increasing the bound to 0.45, produces larger gains for the normal and 1WN, at the expense of decreased efficiency for the slash. Even so, the estimated efficiencies are all greater than or equal to those in Table 3 for $c_{20} = 1.1$. Lower bounds much larger than 0.45, however, seem to cause too much harm on the slash to be an appropriate tuning method. We note that the modification discussed in this paragraph adds only one line of code to the computer program which finds λ^*.

To this point our Monte Carlo study has been limited to sample size n = 20. By the adaptive nature of $\hat{\theta}^*_{(1)}$, however, the sample size should be an important factor in its performance. We have argued previously that as n increases, the performance of the adaptive

TABLE 5

RELATIVE EFFICIENCY OF $\hat{\theta}^*_{(1)}$ TO THE BISQUARE
FOR n = 15 AND n = 40; p = 3.0

	Normal	Wild Normal	Slash
n = 15, $c_{15} = 1.15$	105.2 (0.3)	97.7 (0.4)	100.8 (0.4)
n = 15, $c_{15} = 1.15$, $\frac{1}{n}\Sigma\psi'(\lambda y_i) \geq 0.40$	105.5 (0.3)	98.3 (0.2)	101.0 (0.4)
n = 15, $c_{15} = 1.15$, $\frac{1}{n}\Sigma\psi'(\lambda y_i) \geq 0.45$	106.1 (0.3)	99.1 (0.2)	100.4 (0.3)
n = 40, $c_{40} = 0.80$ (2WN)	106.6 (0.3)	99.8 (0.3)	106.1 (0.6)

estimator should improve relative to that of the nonadaptive one. We illustrate this quantitatively in Table 5 for n = 15 (about the smallest n for which $\hat{\theta}^*_{(1)}$ might prudently be used) and n = 40. For n = 15, the 1WN consists of 14 standard normals and one N(0,100) in every sample. For n = 40, we used a "two wild normal"--38 standard normals and two N(0,100).

Even for n = 15 (and $c_{15} = 1.15$), $\hat{\theta}^*_{(1)}$ performs comparably to the bisquare. Since it is substantially better for the normal and substantially worse for the 1WN, there is no clear-cut winner. Placing a lower bound on $1/n \Sigma \psi'(\lambda y_i)$ produces incremental changes similar to those illustrated in Table 4. Even with those improvements the advantage of $\hat{\theta}^*_{(1)}$ over the bisquare is probably not large enough to warrant its use.

By sample size n = 40, the advantage of the adaptive estimator is certainly significant. While it is comparable to the bisquare for the 2WN, $\hat{\theta}^*_{(1)}$ beats the bisquare by at least 6 percent for both the normal and slash. Thus $\hat{\theta}^*_{(1)}$ quickly approaches the ultimate advantage which is achieved as $n \to \infty$.

11. Conclusions

Perhaps the first conclusion of any robustness study should be the advice--it is not so important which robust technique you use, as long as you do use one. The gains from using a new and improved robust estimator are small compared to the risks of not using one

at all. On the other hand, the search for improvements over existing methods can lead to very important insights.

The scale factor (or its inverse, the scale parameter) is an important determinant of the efficiency of an M-estimator and thus deserves much more attention than it has received in the past. Accurate tuning of the scale factor to the underlying distribution function can lead to very substantial gains in asymptotic efficiencies in comparison to those of nonadaptive M-estimators. By minimizing the estimated variance as a function of the scale factor, one achieves the best possible asymptotic variance using any scale factor. Furthermore, since only one parameter is chosen adaptively, surprising gains are possible for samples as small as 15 or 20. Perhaps this fact has gone unnoticed because of the manipulations in Section 8 which are required to make the estimator acceptable.

Adaptive choice of the scale factor also offers advantages in the reporting stages of an analysis. If F is normal, there is a good chance that λ^* will equal 0, and the mean can be used. While analysts using other robust methods might also realize that the data supports the normal hypothesis, this assessment is likely to depend on a subjective look at the data. On the other hand as λ^* increases, one has more evidence that the normal hypothesis is untenable. Mallows (1979) advocates reporting the weights that each observation has on the analysis. For M-estimation of location, these are $W_i = \psi[\lambda(x_i - \hat{\theta})]/\lambda(x_i - \hat{\theta})$. When λ is chosen adaptively, the set of weights is much more likely to be appropriate.

The success of the adaptive M-estimator of location suggests that the same principle has merit for other robust estimation problems. The most straightforward generalization is to robust linear regression. Residuals from an initial resistant fit would replace $x_i - M$ in the definition of $\hat{V}(\lambda)$. Issues like the effect of p(the number of independent variables) on the small sample behavior and the sensitivity of the estimator to the initial fit will require more investigation. Other potential uses for the principle include robust estimation in general one parameter models (Huber; 1977; p. 32) and in generalized linear models (Pregibon, 1979). In each case the weights of the observations can depend on a parameter which is chosen adaptively.

APPENDIX

A. Program to Compute λ^* and $\hat{\theta}^*_{(1)}$

The FORTRAN subroutine ESTIM uses the algorithm in Section 7 to compute λ^* and $\hat{\theta}^*_{(1)}$ and returns those values as FLMBDA and T2 respectively. ESTIM calls two other subroutines FG, which calculates $\hat{V}(\lambda)$ and $\hat{V}'(\lambda) + g_n(\lambda)$, and PS, which calculates $\psi(\lambda y)$, $\psi'(\lambda y)$ and $\lambda y\, \psi''(\lambda y)$. FG also calls PS.

The following values must be passed to ESTIM:

N	the sample size n
X	the sample in X(1),...,X(N)
Y	order statistics of the absolute differences from the median
FMED	sample median
FMAD	sample MAD
PARAM (1)	c_n
PARAM (2)	tolerance for binary search = PARAM (2)/FMAD PARAM (2) = 0.06 is sufficiently small
PARAM (3)	p for ψ_p; if PARAM (3) $\leq$ 0.5, p = ∞
PARAM (4)	minimum value for $1/n\ \Sigma\ \psi'(\lambda y_i)$; see Table 4, Section 10
IODBUG	if greater than zero, debugging information is written on unit IODBUG at each step of search for λ^*

```
      SUBROUTINE ESTIM (T2,FLMBDA,IODBUG)
      COMMON /PARAM/ PARAM(4)
      COMMON /X/ X(100), Y(100), FMED, FMAD, N
C       A = LOWER BOUND FOR ROOT OF (VHAT' + G)
      A = 0.001/FMAD
      EPS = PARAM(2)/FMAD
      NN = N + 1
C       FA = VHAT(A)
C       GA = VHAT'(A) + G(A)
      CALL FG(A,FA,GA)
      IF (IODBUG.LE.0) GO TO 155
      KK = 155
      WRITE (IODBUG,150) NN,A,FA,GA,KK
  150 FORMAT (I5,F8.4,2F9.3,I6)
C       CHECK FOR ROOT AT ZERO
  155 IF (GA.GE.0.AND.Y(N).LT.100*FMAD) GO TO 210
      NN = N
  160 B = 1.0/Y(NN)
      CALL FG(B,FB,GB)
      IF (IODBUG.LE.0) GO TO 165
      KK = 165
      WRITE (IODBUG,150) NN,B,FB,GB,KK
C       CHECK FOR ROOT BETWEEN A AND B
  165 IF (GB.GE.0) GO TO 180
C
C       GB IS STILL NEGATIVE
      A = B
      GA = GB
      NN = NN - 1
      IF (2*NN.GT.N) GO TO 160
      FLMBDA = 1.0/Y(NN+1)
      GO TO 200
C
C       ROOT BETWEEN  A  AND  B
C       DO BINARY SEARCH UNTIL B-A < EPS
  180 DIFF = B - A
      IF (DIFF.LT.EPS) GO TO 195
      AB = (A+B)/2.0
      CALL FG(AB,FAB,GAB)
      IF (IODBUG.LE.0) GO TO 185
      KK = 185
      WRITE (IODBUG,150) NN,AB,FAB,GAB,KK
  185 IF(GAB.GE.0) GO TO 190
C
C       ROOT BETWEEN  AB  AND  B
      A = AB
      GA = GAB
      GO TO 180
C
C       ROOT BETWEEN  A  AND  AB
  190 B = AB
```

```
      GB = GAB
      GO TO 180
C
C       ROOT BETWEEN  A  AND  B;  B-A < EPS
  195 FLMBDA = B - DIFF*GB/(GB-GA)
  200 SUM1 = 0.0
      SUM2 = 0.0
      DO 205 J=1,N
      Z = ( X(J) - FMED ) * FLMBDA
      CALL PS(Z,PSI,PSI1,ZPSI2)
      SUM1 = SUM1 + PSI
      SUM2 = SUM2 + PSI1
  205 CONTINUE
      T2 = FMED + SUM1/(SUM2*FLMBDA)
      RETURN
C
C       ROOT AT ZERO
C       T2 = SAMPLE MEAN
  210 FLMBDA = 0
      T2 = 0
      DO 215 J=1,N
      T2 = T2 + X(J)
  215 CONTINUE
      T2 = T2/N
      RETURN
      END


      SUBROUTINE FG(A,FA,GA)
      DIMENSION SUM(5)
      COMMON /X/ X(100),Y(100),FMED,FMAD,N
      COMMON /PARAM/ PARAM(4)
      CN = PARAM(1)
      DO 120 M=2,5
      SUM(M) = 0.
  120 CONTINUE
      DO 140 J=1,N
      Z = (X(J) - FMED) * A
      CALL PS (Z,PSI,PSI1,ZPSI2)
      PSISQ = PSI*PSI
      ZPPSI1 = Z*PSI*PSI1
      SUM(2) = SUM(2) + PSISQ
      SUM(3) = SUM(3) + PSI1
      SUM(4) = SUM(4) + ZPSI2 - Z*Z*PSISQ*CN
      SUM(5) = SUM(5) + ZPPSI1
  140 CONTINUE
      FA = N*SUM(2) / (A*A*SUM(3)*SUM(3))
      FACTOR = 2*N / (A*A*A*SUM(3)*SUM(3) )
      GA = FACTOR * ( SUM(5) - SUM(2) - SUM(4)*SUM(2)/SUM(3) )
      IF (SUM(3).LT.N*PARAM(4)) GA = 10000.
      RETURN
      END
```

```
      SUBROUTINE PS(Z,PSI,PSI1,ZPSI2)
      COMMON /PARAM/ PARAM(4)
      P = PARAM(3)
      ZSQ = Z*Z
      IF (P.GT.0.5) GO TO 5
C
C       P = INFINITY
      IF (ABS(Z).GT.8.) GO TO 10
      EX = EXP (-0.5*ZSQ)
      PSI = Z*EX
      PSI1 = (1.-ZSQ)*EX
      ZPSI2 = ZSQ*(ZSQ-3.0)*EX
      RETURN
C
C       P > 0.5
    5 CONTINUE
      IF (ABS(Z).GT.1000) GO TO 10
      P2 = 2.0*P - 1.0
      DZSQ1 = 1.0 + ZSQ/P2
      DZSQ1P = DZSQ1**P
      PSI = Z/DZSQ1P
      PSI1 = (1.0-ZSQ) / (DZSQ1P*DZSQ1)
      ZPSI2 = -2.0*ZSQ*P*(3.0-ZSQ) / (P2*DZSQ1P*DZSQ1*DZSQ1)
      RETURN
   10 PSI = 0
      PSI1 = 0
      ZPSI2 = 0
      RETURN
      END
```

APPENDIX

B. Technical Details of the Monte Carlo Study

In Section 10 we present Monte Carlo results comparing various forms of the adaptive M-estimator with the nonadaptive bisquare. In this appendix we give details of how these results were obtained. Most of the numbers presented in the text are relative efficiencies. By using two well-known variance reduction techniques, the standard errors of these numbers have been greatly reduced from what they otherwise would be. The first technique is the Monte Carlo swindle for location estimation, and the second is the use of common streams of pseudo random numbers for comparison of correlated estimators; see, e.g., Kleijnen (1975).

The relative efficiency of the estimator T_2 to T_1 is the ratio $E(T_1 - \theta)^2 / E(T_2 - \theta)^2$. We estimate this by the ratio of estimated expected squared errors. The expected squared error of T_1 may be estimated naively by $1/N \sum_{j=1}^{N} [T_1(\underset{\sim}{x}^{(j)}) - \theta]^2$, the mean of squared errors for the N pseudo random samples. A great reduction from the variance of this estimate is made possible by the Monte Carlo swindle used in Andrews et al. (1972; Section 4B) and explained in detail by Simon (1976). The simplest example of the swindle is for normal samples. Since $\bar{X}$ is the Pitman estimator and T_1 is equivariant, $E[T_1(\underset{\sim}{X}) - \theta]^2 = E[T_1(\underset{\sim}{X}) - \bar{X}]^2 + E(\bar{X} - \theta)^2 = E[T_1(\underset{\sim}{X}) - \bar{X}]^2 + 1/n$. Because the variance of $[T_1(\underset{\sim}{X}) - \bar{X}]^2$ is usually many times smaller than that of $[T_1(\underset{\sim}{X}) - \theta]^2$, the swindle is very useful. The general swindle

reduces the variance whenever X can be factored into a normal divided by an independent random variable. When T_1 was the bisquare and $n = 20$, the swindle reduced the variance of the estimate of $E[T_1(\underset{\sim}{X}) - \theta]^2$ by factors of approximately 25, 32, and 2.3 for the normal, 1WN and slash respectively. Since this evaluation of the swindle requires estimation of $E[T_1(\underset{\sim}{X}) - \theta]^4$, we receive a free estimate of the kurtosis of the bisquare. For the normal and 1WN the estimated kurtosis is not significantly different from zero. For the slash, however, the estimated kurtosis of the bisquare is 0.8.

The second variance reduction technique uses the fact that estimates of $E[T_1(\underset{\sim}{X}) - \theta]^2$ and $E[T_2(\underset{\sim}{X}) - \theta]^2$ derived from the same sequence of samples are likely to be highly correlated. This reduces the variance of their ratio, the estimated relative efficiency of T_2 to T_1, in comparison with the ratio based on different samples for T_1 and T_2. The results in Table 1 of Section 10 are based on 10,000, 20,000, and 100,000 samples from the normal, 1WN and slash respectively. In order to obtain an estimate of the standard error of the estimated relative efficiency, the entire Monte Carlo was divided into 100 equal parts. For each sub-Monte Carlo of 100, 200, or 1000 samples an estimated relative efficiency was calculated. Since the estimated relative efficiency of T_2 to T_1 is approximately the mean of the 100 separate relative efficiencies, 1/10 times the standard deviation of these relative efficiencies provides an estimate of the standard error of the overall estimated relative efficiency.

Achieving the precision of the estimated relative efficiencies in Table 1 required a sizeable Monte Carlo study. A total of 12 hours on a PDP11/34A was required to estimate the triefficiency of $\hat{\theta}^*_{(1)}$. In order to reduce the computations necessary to obtain comparable precision of the triefficiencies for the ten variations of $\hat{\theta}^*_{(1)}$ appearing in Tables 2-4, we employed an additional variance reduction technique.

Suppose that T_3 is a small modification of T_2. Then the performance of T_3 relative to T_1 will be correlated with the performance of T_2 relative to T_1. By reusing a fraction of the 100 sets of samples, we may take advantage of this correlation to reduce the variance of the estimate. In the following derivation, $m = 20$, $M = 100$, and (X_i, Y_i) are the relative efficiencies of T_2 to T_1 and T_3 to T_1 from the i-th set of samples. The number of samples in each sub-Monte Carlo is large enough to insure approximate normality of X_i and Y_i.

Suppose that we observe $(x_1,y_2),\ldots,(x_m,y_m)$ and $x_{m+1},\ldots,x_M$ where

$$\text{(B.1)} \qquad \begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \right)$$

are independent for $i = 1,\ldots,M$. We wish to estimate μ_x and μ_y. While $\bar{x}_M = \frac{1}{M}\Sigma_{i=1}^{M} x_i$ is the optimal estimate of μ_x, $\bar{y}_m = \frac{1}{m}\Sigma_{i=1}^{m} y_i$ is not necessarily the best estimator of μ_y.

The conditional distribution of Y given X is

$$\text{(B.2)} \qquad Y|X=x \sim N(\alpha+\beta x,\ (1-\rho^2)\sigma_y^2)$$

where $\alpha = \mu_y - \rho \frac{\sigma_y}{\sigma_x} \mu_x$ and $\beta = \rho \frac{\sigma_y}{\sigma_x}$. Anderson (1957) showed that the maximum likelihood estimate of μ_y is $\hat{\mu}_y = \hat{\alpha} + \hat{\beta}\bar{x}_M$ where $\hat{\alpha}$ and $\hat{\beta}$ are the least squares estimators from a regression of y on x. Thus

$$\hat{\mu}_y = \bar{y}_m - \hat{\beta}(\bar{x}_m - \bar{x}_M) \tag{B.3}$$

and

$$\begin{aligned} \mathrm{Var}(\hat{\mu}_y) &= \mathrm{Var}_{\underset{\sim}{x}} E(\hat{\mu}_y | \underset{\sim}{x}) + E_{\underset{\sim}{x}} \mathrm{Var}(\hat{\mu}_y | \underset{\sim}{x}) \\ &= \mathrm{Var}(\mu_y - \rho \frac{\sigma_y}{\sigma_x} \mu_x + \rho \frac{\sigma_y}{\sigma_x} \bar{x}_M) \\ &\quad + E(1-\rho^2)\sigma_y^2 \left[\frac{1}{m} + \frac{(\bar{x}_m - \bar{x}_M)^2}{\sum_{i=1}^{m} (x_i - \bar{x}_m)^2} \right] \\ &= \rho^2 \frac{\sigma_y^2}{M} + (1-\rho^2) \frac{\sigma_y^2}{m} + (1-\rho^2)\sigma_y^2 \, E \frac{(\bar{x}_m - \bar{x}_M)^2}{\sum_{i=1}^{m} (x_i - \bar{x}_m)^2}. \end{aligned} \tag{B.4}$$

Suppose that $\{x_1,\ldots,x_m\}$ is a random subset of $\{x_1,\ldots,x_M\}$; then for large M, $\bar{x}_M \to \mu_x$, and $(\bar{x}_m - \bar{x}_M)^2 / \Sigma_{i=1}^{m} (x_i - \bar{x}_m)^2$ is approximately $1/m(m-1)$ times a random variable distributed as F with 1 and $(m-1)$ degrees of freedom. However, $\{x_1,\ldots,x_m\}$ need not be a random subset of $\{x_1,\ldots,x_M\}$. By the nature of the Monte Carlo procedure we may generate $\{x_1,\ldots,x_M\}$ and then $Y_i | x_i$ for any values of i that we choose. It makes sense then to choose $\{x_1,\ldots,x_m\}$ to minimize $(\bar{x}_m - \bar{x}_M)^2 / \Sigma_{i=1}^{m} (x_i - \bar{x}_m)^2$. In practice, for moderately large m and M, the last term can be made negligible. Thus we have

(B.5) $$\mathrm{Var}(\hat{\mu}_y) \approx \rho^2 \frac{\sigma_y^2}{M} + (1-\rho^2) \frac{\sigma_y^2}{m} ,$$

so that $\mathrm{Var}(\hat{\mu}_y)$ is virtually a convex combination of the variance from using m and the variance from using M observations. The larger ρ^2 is, the greater the efficiency of the procedure.

The estimates of ρ^2 for the estimators in Tables 2-4 range from 0 to 0.99. Two estimators ($c_{20} = 0.0$ and the adaptive ψ_{bs}) are practically uncorrelated with the basic adaptive estimator. In all other cases except two ($\psi_{1.5}$ on the 1WN and $c_{20} = 1.3$ on the 1WN) the estimate of ρ^2 is at least 0.50, and in over half of these it exceeds 0.85. When $\rho^2 = 0.85$, the variance from reusing 20 sets of samples (such that $(\bar{x}_m - \bar{x}_M) = 0$) is the same as that from 62.5 new sets of samples. Thus 62.5 percent of the information is available from just 20 percent of the samples.

There is another important benefit of reusing samples. The main purpose of Tables 2-4 is to demonstrate the effects of various parameters on the performance of $\hat{\theta}^*_{(1)}$. Because of the high positive correlations among entries in the same columns (even across tables), the standard error of the difference between two estimated relative efficiencies is usually smaller than either standard error. Of course, this phenomenon is most pronounced for very similar estimators.

References

Anderson, T.W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. J. Amer. Statist. Assoc., 52, 200-203.

Andrews, D.F. (1974). A robust method for multiple linear regression. Technometrics, 16, 523-531.

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972). Robust Estimates of Location. Princeton University Press.

Beran, R. (1974). Asymptotically efficient adaptive rank estimates in location models. Ann. Statist., 2, 63-74.

Beran, R. (1977). Robust location estimates. Ann. Statist., 5, 431-444.

Berk, R. (1967). A special structure and equivariant estimation. Ann. Math. Statist., 38, 1436-1445.

Gross, A.M. (1976). Confidence interval robustness with long-tailed symmetric distributions. J. Amer. Statist. Assoc., 71, 409-416.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. J. Amer. Statist. Assoc., 69, 383-393.

Hodges, J.L., Jr. and Lehmann, E.L. (1963). Estimates of location based on rank tests. Ann. Math. Statist., 34, 598-611.

Hogg, R.V. (1974). Adaptive robust procedures. J. Amer. Statist. Assoc., 69, 909-927.

Huber, P.J. (1964). Robust estimation of a location parameter. Ann. Math. Statist., 35, 73-101.

Huber, P.J. (1972). Robust statistics: A review. Ann. Math. Statist., 43, 1041-1067.

Huber, P.J. (1977). Robust Statistical Procedures. Society for Industrial and Applied Mathematics, Philadelphia.

Jaeckel, L.A. (1971). Some flexible estimates of location. Ann. Math. Statist., 42, 1540-1552.

Johns, M.V. (1974). Nonparametric estimators of location. J. Amer. Statist. Assoc., 69, 453-460.

Johns, M.V. (1979). Robust Pitman-like estimators. In Robustness in Statistics, 49-60, eds., Launer, R.L. and Wilkinson, G.N. Academic Press, New York.

Kleijnen, J.P.C. (1975). Statistical Techniques in Simulation, Vol. I and II. Marcel Dekker, New York.

Mallows, C.L. (1979). Robust methods--Some examples of their use. Amer. Statistician, 33, 179-184.

Pregibon, D. (1979). Data analytic methods for generalized linear models. Doctoral thesis, University of Toronto.

Sacks, J. (1975). An asymptotically efficient sequence of estimators of a location parameter. Ann. Statist., 3, 285-298.

Simon, G. (1976). Computer simulation swindles, with applications to estimates of location and dispersion. Appl. Statist., 25, 266-274.

Stein, C. (1956). Efficient nonparametric testing and estimation. Proc. Third Berkeley Symposium on Math. Statist. and Prob., 1, 187-196.

Stone, C.J. (1975). Adaptive maximum likelihood estimators of a location parameter. Ann. Statist., 3, 267-284.

Switzer, P. (1971). Efficiency robustness of estimators. Proc. Sixth Berkeley Symposium on Math. Statist. and Prob., 1, 283-291.

Takeuchi, K. (1971). A uniformly asymptotically efficient estimator of a location parameter. J. Amer. Statist. Assoc., 66, 292-301.

Tukey, J.W. (1979). Study of robustness by simulation. In Robustness in Statistics, 75-102, eds., Launer, R.L. and Wilkinson, G.N. Academic Press, New York.

Van Eeden, C. (1970). Efficiency-robust estimation of location. Ann. Math. Statist., 41, 172-181.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER No. 3	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* AN ADAPTIVE CHOICE OF THE SCALE PARAMETER FOR M-ESTIMATORS		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Robert Michael Bell		8. CONTRACT OR GRANT NUMBER(s) ARO DAAG29-79-C-0166
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, California		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P. O. Box 12211 Research Triangle Park, NC 27709		12. REPORT DATE July 1, 1980 13. NUMBER OF PAGES 72
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Robust Estimation of Location, Adaptive Estimation, M-Estimate, Tri-Efficiency, Asymptotic Efficiency, Asymptotic Normality, Scale Parameter for M-Estimators.

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Let $x_1,\ldots,x_n$ be a random sample from a distribution symmetric about the unknown location parameter θ. A major class of robust estimators of location is the class of M-estimators, each of which corresponds to a function ψ defined on the reals. For a given function ψ, the variance of the corresponding M-estimator varies considerably with the value of the scale parameter. It is therefore proposed that the value which minimizes an estimate of the asymptotic variance of the M-estimator be used as the scaling factor. The performance of the estimator for small samples is investigated by Monte Carlo methods for choices of ψ.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*